A Unified Framework for Finite-Mask Coronagraphy as Applied to Exoplanet Imaging

by

David Gordon Ely

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

September, 2018

Abstract

A large number of different coronagraphic strategies have been proposed for the purpose of directly imaging recently discovered exoplanets. To date, these methods have depended on analysis of numerical simulations for theoretical understanding of the system responses to illumination under different conditions. We demonstrate here that a key mathematical principle exists, which underlies all finite-mask coronagraphy. We use this to develop the naturally propagating modes from pupil plane to Lyot plane with explicit reference to perturbations, both from plane waves and central design wavelength, while retaining mostly arbitrary mask behavior. The computational work necessary for this method is comparable to current one-dimensional approaches.

Thesis Committee

Primary Readers

Laurent Pueyo (Primary Advisor)
Associate Astronomer
Space Telescope Science Institute

Julian Krolik (Advisor of Record)
Professor
Rowland Department of Physics and Astronomy
Johns Hopkins Krieger School of Arts and Sciences

Jared Kaplan
Associate Professor
Rowland Department of Physics and Astronomy
Johns Hopkins Krieger School of Arts and Sciences

Anand Sivaramakrishnan
Observatory Scientist
Space Telescope Science Institute

Rémi Soummer
Associate Astronomer with Tenure
Space Telescope Science Institute

Acknowledgments

I would like to thank Laurent, for his patience and advice; Julian, for the same, and for the consistent TA position; Colin Norman, for suggesting this line of investigation; my family and friends, for their support; and of course, all others who have aided me in my time at Johns Hopkins.

For Dad.

Table of Contents

Table of Contents v

List of Tables ix

List of Figures x

1 Introduction 1

2 Mathematics of the Slepian problem 14

2.1 Initial Setup 16

2.2 Change of notation 20

2.3 New notation with the Slepian problem 24

2.4 Abstract kernel properties 26

2.5 Dual kernel 29

2.6 Basis considerations 33

2.7 Summary 35

3 Application to the General APLC 38

3.1 Instrument Layout and Notation 41

3.1.1 Pupil plane . . . 41
3.1.2 Image plane . . . 43
3.1.3 Lyot and Instrument planes . . . 44
3.1.4 Non-dimensional coordinates . . . 45
3.1.5 Abstract expressions . . . 50
3.2 Optical Reversal for Finite Mask Applications . . . 54
3.3 Pupil-Mask Basis Functions and their Immediate Applications 58
3.3.1 Choice and Implications of $t \leq T$ . . . 62
3.3.2 Restrictions on (t,m) . . . 66
3.3.3 Simple operators on basis functions . . . 74
3.3.4 Basis functions for rectangular and other masks . . . 82
3.4 Algorithm for determining Slepian functions . . . 84
3.5 Summary . . . 86

4 Propagation with Slepian Modes **91**

4.1 Monochromatic propagation . . . 93
4.1.1 Single Slepian apodization . . . 95
4.1.2 Combined Slepian apodization . . . 104
4.2 Broadband behavior . . . 110
4.2.1 The Dilation Operator . . . 110
4.2.2 Abstract application of the Dilation Operator . . . 113
4.2.3 Direct check of dilation operator performance . . . 117

4.2.4 Broadband propagation 122
4.3 Perturbations and off-axis plane waves 129
4.3.1 Pupil-limited operators as linear operators 129
4.3.2 Polynomial perturbations 133
4.3.3 Off-axis plane waves 136
4.3.4 Propagation of perturbations and off-axis plane waves 147
4.4 Summary 153

5 Circular Pupil Slepians and Non-Circular Demonstration 161
5.1 Circular APLC: Prior and New Results 163
5.1.1 Analytical Simplifications for Circular Pupils 163
5.1.2 Comparison to Prior Results 169
5.1.3 Angular modes 171
5.1.4 Bell-bagel transition 177
5.2 Non-Circular Demonstration: JWST as APLC 181
5.2.1 Coronagraphic set-up 181
5.2.2 Apodization and PSF results 185
5.2.3 Instrument Plane Response to Combined Apodizations 192
5.2.4 Eigensystem Trends with Changing Wavelength 196
5.3 Conclusions 207
5.3.1 Circular 207
5.3.2 Non-Circular 208

6 Conclusion 212

7 Curriculum Vitae 225

List of Tables

2.1 New notation 21

3.1 Nondimensional Distances 47

3.2 Coronagraphic Slepian notation 53

3.3 Mismatched phase mask integrals 71

4.1 Mask expansions 94

4.2 Primary Seidel aberrations 135

5.1 JWST metrics 191

List of Figures

1.1 Exoplanet chart 2

1.2 Notable first direct images of exoplanets. 5

2.1 Slepian Problem illustrated 15

2.2 Diverging eigenfunctions 33

3.1 APLC Layout 41

3.2 Pupil geometry 42

3.3 Coronagraph coordinates layout 48

3.4 Inscription of pupil 50

3.5 P_3, P_1 Venn diagram 52

3.6 Pupil plane basis functions 60

3.7 Image plane basis functions 61

3.8 Number of basis functions 63

3.9 Kernel Element Ratio 65

3.10 Kernel Element Values 65

3.11 Cutoff changes Λ_a 66

3.12 Λ_a by cutoff 67
3.13 Λ_a by cutoff, 2 67
3.14 Convergence of "mismatched" basis functions 73
3.15 Convergence of $r^n \left|tm\right\rangle$ 79
3.16 Convergence of $\mathcal{J}_{t+1}(r) \times \mathcal{J}_{s+1}(r)$ 80

4.1 Coronagraph coordinates layout repeated 92
4.2 Basis function dilation 118
4.3 Dilated eigenvalues 120
4.4 Dilated eigenvalues, different T 121
4.5 Constant pupil reproduction 133
4.6 Constant pupil reproduction 133
4.7 Reproducing r^n with direct matrix vs. $1/\Lambda$ 135
4.8 Off-axis, $\omega = 0.1$ 139
4.9 Relative error in off-axis, $\omega = 0.1$ 140
4.10 Off-axis, $\omega = 1.0$ 141
4.11 Relative error in off-axis, $\omega = 1.0$ 142
4.12 Off-axis, $\omega = 5.0$ 143
4.13 Relative error in off-axis, $\omega = 5.0$ 144
4.14 Relative error in off-axis, $\omega = 5.0$ 146

5.1 Comparison of residuals. 169
5.2 Comparison of throughputs. 170

5.3 Angular modes 171
5.4 Throughput, residual for $|m| = 0$ 172
5.5 Throughput, residual for second $|m| = 0$ 172
5.6 Throughput, residual for $|m| = 1$ 173
5.7 Throughput, residual for second $|m| = 1$ 173
5.8 Throughput, residual for $|m| = 2$ 174
5.9 Throughput, residual for second $|m| = 2$ 174
5.10 Overlay of residuals 175
5.11 Residuals at different $\mathcal{N}$ 175
5.12 Residuals at different $\mathcal{N}$, 2 176
5.13 Eigenvalues vs. prediction 177
5.14 Bell-bagel transition 178
5.15 Bell-bagel transition contours 179
5.16 Changing relevance of b_i 180
5.17 JWST mirror 182
5.18 JWST kernel 185
5.19 JWST eigenvalues 185
5.20 JWST pupil, PSF 187
5.21 JWST eigenvalues 188
5.22 JWST eigenvalues 189
5.23 JWST radial profile reduction 190
5.24 JWST radial profiles 190

5.25 JWST near-circular apod. 193
5.26 JWST circular radial contrast 193
5.27 JWST circular radial contrast 195
5.28 Shannon number comparison 197
5.29 Shannon number comparison 197
5.30 Top eigenvalues 199
5.31 Top eigenvalues 199
5.32 Top residuals 200
5.33 Apod. ordering shift 200
5.34 Top eigenvalue derivatives 201
5.35 Eigenvalue derivatives 2 202
5.36 Eigenvalue derivatives 3 202
5.37 Broadband apodizations 1 203
5.38 Broadband apodizations 2 204
5.39 JWST Λ_a, fixed $\mathcal{N}$ 205
5.40 Trends in Λ_a fits 206

Chapter 1

Introduction

The physical properties of star-orbiting exoplanets is one of the most engaging fields currently being studied. While an exoplanet around another astronomical body was first recognized in 1992 (Wolszczan and Frail, 1992), the first confirmed discovery around a main-sequence star was at 51 Pegasi in 1995 by Mayor and Queloz (Mayor and Queloz, 1995).

The method used for this original detection was the detection of the Doppler shift in the star's emissions caused by the revolution around the barycenter of the star-planet system. The stark contrast of this and later massive, short-orbit "hot-Jupiters" to our own Solar System has required considerable research into the mechanics of planetary system formation. As of July 2018, 3801 exoplanetary discoveries have been confirmed (*The Extrasolar Planets Encyclopaedia*). Figure 1.1 shows a sample of these, sorted by planetary mass and orbital period.

Since those first discoveries, a range of different techniques have been developed for detection, mostly reliant on indirect approaches. The radial

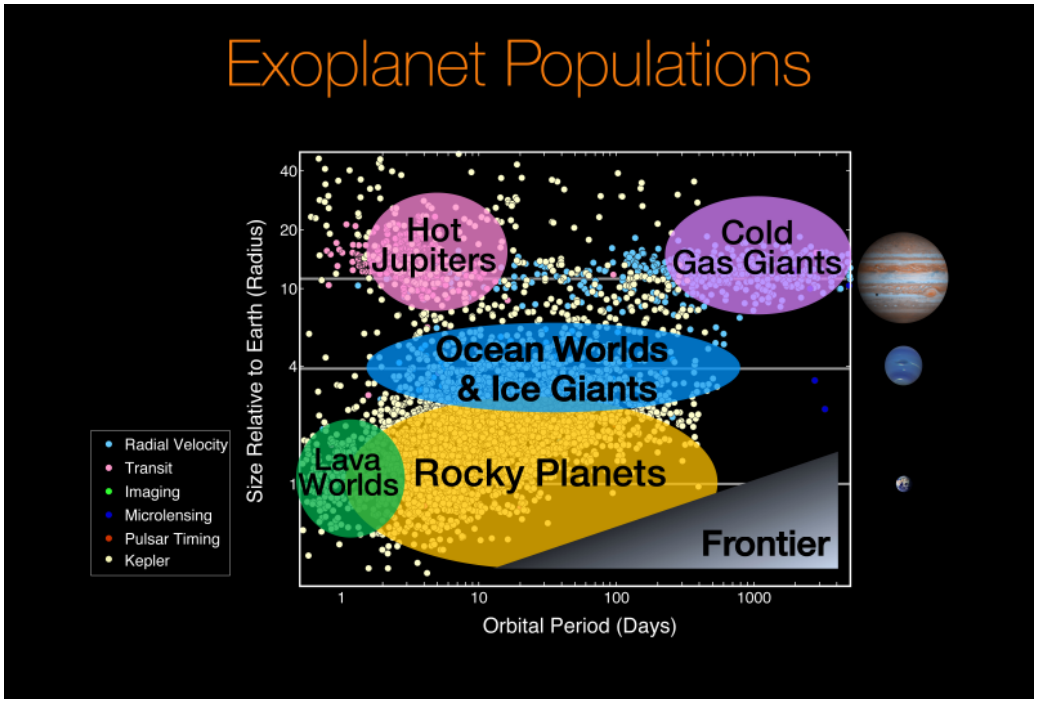


Figure 1.1: Chart of exoplanets as of June 2017. Picture credit NASA/Ames Research Center, produced by Natalie Batalha and Wendy Stenzel under public domain. (*Exoplanet Populations*)

velocity method, mentioned above, naturally favors discovery of heavy planets close to their star. As this method can be used with any well-detectable spectral feature, any spectroscopic observatory is equipped to make these measurements. Current observations (*The Extrasolar Planets Encyclopaedia*) (Burrowsa and Marcyb, 2014, ch. 9) have clustered around one to ten Jupiter masses, with orbital periods of 100 to 4000 days a few hundred days.

Direct-transit searches measure the decrease in stellar luminosity due to obstruction by the exoplanet (primary transit) or by the decrease caused from loss of reflected light when the planet is eclipsed (secondary transit). This method's pre-Kepler detections (Burrowsa and Marcyb, 2014, ch. 9) cluster around masses one-third to ten times Jupiter's, with one to ten day orbits;

Kepler itself predominately discovered planets of sizes 0.8 to four times Earth's radius (Burrowsa and Marcyb, 2014, ch. 9).

Direct-transit detections began in 1999 (*Exoplanet Exploration: Historic Timeline*) with the detection at HD 209458 (Charbonneau et al., 2000), though the first planetary discovery did not occur until the 2002 OGLE-III detection at OGLE-TR-56 (Udalski et al., 2002). The dedicated Kepler space telescope and subsequent K2 run observed over 2600 confirmed exoplanets so far (*Kepler and K2*) over 18 missions from launch in 2009 to July of 2018. It uses wavelengths from 400 to 850 nm to do so (*Kepler and K2*). TESS, launched in June of 2018 (*Transiting Exoplanet Survey Satellite*) with four 0.1m apertures and sensitive in the 600-1000 nm range, is expected to be just as successful detecting exoplanets in nearby and brighter stars than Kepler (*TESS NasaFacts*). The JWST (*James Webb Space Telescope*), currently planned for launch in 2021, will spend a portion of its time on these observations. It has a 3.5m segmented main mirror and will be sensitive to wavelengths between .6 and 28 microns. From the ESA, the small CHEOPS (*CHaracterising ExOPlanet Satellite*) will have four cameras for working with 400-1100nm light. It is intended to launch in 2019 (*The CHEOPS Mission*)

While other methods exist, the one of interest to us in this thesis is the prospect of direct imaging of the planets in question. While the pleasure of viewing such a picture is appealing, it is the scientific data available from such observations which make these directions worthwhile. The spectra from these observations provide information on the atmospheric composition, planetary composition, and atmospheric dynamics, among others (Traub and

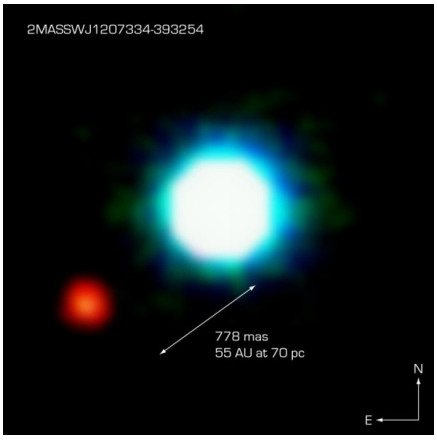
Oppenheimer, 2010). Sufficiently large space telescopes might resolve the planet across multiple pixels, offering our first look out onto the surface of foreign territories.

The ability to directly resolve the planet requires two main goals that we will focus on in this thesis. The first is a sufficient extinction of stellar light to suppress the planet/star contrast. The second is the innermost working angle at which that contrast can be achieved. These must be weighed in any real design against all other factors, including throughput of light, engineering capabilities, etc.

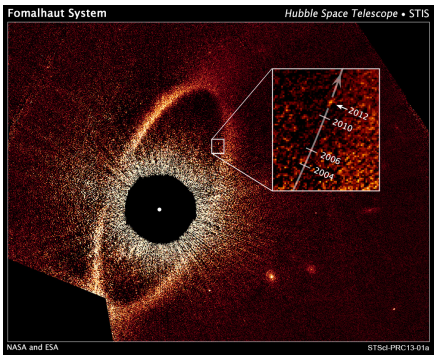
Direct imaging began with 2M1207b orbiting a brown dwarf, observed in the IR by the VLT in 2004 (*2M1207b - First image of an exoplanet*). The Spitzer Space Telescope (*Spitzer Space Telescope*) was launched in August 2003 and sensitive to several bands in the IR from 2 to 200 microns. It detected the first exoplanetary light orbiting a main-sequence star directly in March 2005 (*Exoplanet Exploration: Historic Timeline*), though this light was not imaged (*NASA's Spitzer Marks Beginning of New Age of Planetary Science*). The first visible range (600nm), star-orbiting image was made with the Hubble's 2.5m primary mirror. The picture, of Fomalhaut b, was released in November 2005 (*Hubble Directly Observes Planet Orbiting Fomalhaut*).

WFIRST (*Wide-Field Infrared Survey Telescope*), nearly cancelled in 2018 (*Planetary science wins big in NASA's new spending plan*), will spend a portion of its time dedicated to these images in the range 430-980nm (*WFIRST: Exoplanets - Direct Imaging*). The design goals are to reach 10^{-9} contrast levels at 0.2 arcsecond separation between planet and star using its 2.4m mirror. The JWST

(*James Webb Space Telescope*) will also be capable of direct imaging.



(a) 2M1207b, captured by the VLT in the IR. Credit ESO



(b) Formalhaut b, the first visible-spectrum direct-image exoplanet. This picture is a composite including later observations. Credit NASA and ESA.

Figure 1.2: Notable first direct images of exoplanets.

The planetary light which we wish to capture comes from one of three sources (Traub and Oppenheimer, 2010). Young planets which have recently formed from the planetary disk will still retain tremendous heat, and thus radiate blackbody radiation, usually in the IR and visible bands. Older planets will no longer radiate their own internal heat, but will come to thermal equilibrium with the stellar radiation and so still emit IR blackbody radiation. This radiation will be cooler than the young-planet emissions, and can be modified by greenhouse gas effects.

The other light we can see is reflected stellar radiation, and is therefore dependent on the star's own spectrum and the composition of the planet. Visible light, IR and possibly near UVA light will come from this, thought the last has wavelengths and intensities too small to be useful.

Thus, direct imaging systems' design will focus on wavelengths from 0.4

to ~ 5 microns. A telescope one meter in diameter, looking at a median one micron source, will therefore have a Rayleigh-criteria ($1.22\lambda/D$) diffraction limit of 0.25 mas.

By definition, one parsec is the distance which produces an angular measure of one arc-second for a size of one AU. This means that our .25 mas hypothetical telescope would be able to resolve the Earth at 4 parsecs (13 lightyears). Since the Rayleigh criteria understates the difficulties involved, a more realistic estimate would be 10 parsecs (.63 mas). Jupiter, by contrast, would have a separation of 3.3 mas.

Constrast levels, which do not fall off with difference, can be modeled for different planetary arrangements. The Earth/Sun ratio, for instance, works out as 2.1×10^{-10}, while the Jupiter/Sun ratio is 1.4×10^{-9} (*WFIRST: Exoplanets - Direct Imaging*).

Models of planetary systems around nearby stars, with properties randomly drawn from the current distribution of known exoplanets, indicate that these two regions of parameter space should have a large number of imageable planets. Additionally, the differences in number with contrast changing from 10^{-9} to 10^{-10}, or with inner working angle changing from 2 mas to 1 mas, are substantial (*WFIRST: Exoplanets - Direct Imaging*) (*Exoplanet Probe to Medium Scale Direct Imaging Mission Requirements and Characteristics - (SAG9) Final report* 2015). Correspondingly, we have a great pressure to carefully optimize telescope designs.

There are a variety of different methods which have been created to meet these two challenges. Generically, these are referred to as coronagraph designs,

after Lyot's pioneering work in 1939 (Lyot, 1939) studying the Sun's corona. A tremendous number of different designs since then have been proposed to meet the criteria necessary for exoplanet imaging; we will here briefly discuss some of the different approaches which follow Lyot's general design. (Guyon et al., 2006) has a list of proposals as of 2006.

Lyot included what has become known as the Lyot stop. By introducing a second lens behind the first image plane, he was able to produce a new, second pupil plane. In this plane, he added an annular opening, shaped similarly to the pupil (or, in some cases, undersized along the outside.) This allows the removal of the Airy rings and central spot, significantly improving performance. Many designs now incorporate such a stop, with additional features in the coronagraph.

The first main piece of a coronagraph is a *mask*, a piece of material inside the telescope and on the direct optical path, in an image plane. (Masks which precede the pupil are called *occulters*, first proposed by Lyman Spitzer in 1962 (Spitzer, 1962).) Different shapes have been proposed (circular, square, annular, etc.) to handle the diffraction by diverting a portion of the star's light onto another absorbing surface, producing a specific point-spread function (PSF), or to destructively interfere the starlight with itself.

The mask types have been designed to provide to a variety of phase and amplitude effects. Lyot (Lyot, 1939) built the first successful solid mask coronagraph (Lyot, 1939). The simple nature of the mask, and the hope to improve its low performance at small inner working angles (Guyon et al., 2006), has led to more exotic designs. Multiple Lyot stages can reduce

starlight significantly, at the cost of also suppressing planetary light. The band-limited mask, proposed by Kuchner and Traub (Kuchner and Traub, 2002), is a grayscale mask intended to direct light onto a Lyot stop.

More exotic, and more engineering-difficult, designs have been proposed to affect the phase of the light. These designs can, in theory, supply contrast down to $1.0\lambda/D$, which we have seen opens a highly productive part of parameter space for exoplanet discovery. However, introducing such a phase shift relies on wavelength-dependent techniques, limiting broadband utility.

(Roddier and Roddier, 1997) proposed a mask that would induce a π phase-shift, which is capable of theoretically producing complete destructive interference. The four-quadrant phase mask (Rouan et al., 2000) has two open holes and two π shifts inside a solid mask. This design which is being incorporated into the JWST (*James Webb Space Telescope*). The vortex mask, which induces an angularly-dependent phase change of even order, was suggested in this context by (Foo, Palacios, and Swartzlander, 2005). It is not a small focal-plane mask but rather extends over that entire plane, and can avoid removing planetary light.

The second main piece of the telescope of influence to the coronagraph is the pupil. Circular pupils, the simplest design, produce Airy rings which can easily ruin desired contrast levels. Shaped pupils cause different diffraction patterns, which can result in regions of very improved contrast.

The pupil lens can, itself, selectively alter the amplitude or phase of the incoming light. Such a change is called *apodization* (Jacquinot and Roizen-Dossier, 1964). Commonly, apodization refers to the amplitude effects; if

phase effects are included, they will be referred to explicitly.

The use of pupil shaping to cause effective apodization was started by (Kasdin et al., 2003). Combining apodization and Lyot stops, known as APLC, was researched by (Aime, Soummer, and Ferrari, 2002) (Soummer, 2005). It offered a simple approach based on known methods. (Guyon et al., 2005) proposed a different method of inducing the amplitude shift. Rather than tinting the lens, he showed how a well-designed mirror would induce small phase shifts at the lens, which through small-phase coupling cause amplitude shifts. This method is called Phase Induced Amplitude Apodization, PIAA.

Current methods of generating pupil apodizations produce one of a number of mathematically possible modes which all obey a specific coronagraphic equation (Aime, Soummer, and Ferrari, 2002). While other modes are known to exist, the current methods are poor at finding them. Moreover, analytic methods to study of perturbations to the wavefront are either available, relying on pixellation of the pupil, or require special symmetries. It is precisely these limitations that this thesis will seek to address.

Chapter 2 of this thesis will motivate, and then explain the fundamental mathematical properties of the general type of problem posed by this design. In doing so, it introduces the use of a different mathematical formalism which simplifies some aspects of the analysis.

Chapter 3 will then demonstrate in more detail how this problem and formalism apply to a finite-mask coronagraph. The consideration of mathematically abstract properties allows us to draw a number of important general conclusions. It also will be used to derive simple methods for calculating the

families of apodizations, independent of pupil shape. We will show (in § 4.3) that the empty pupil can be described using the full family of functions. The simple solid and phase masks serve to confirm correspondence of expressions and understanding of equations.

Chapter 4 reveals that in our framework, despite initially only designed for determining apodizations, is capable of handling several additional coronagraphic challenges. Propagation of light from the pupil to the Lyot plane is shown to be naturally described in these modes, with simplifying expressions for several quantities of interest. The dependence of the apodizations on wavelength is shown to have a simple matrix expression in our framework, as does the propagation of light whose wavelength does not match the design wavelength. We are also able to, in principle, accommodate non-planar and off-axis wavefronts. The practical numerical limits restrict these calculations to slowly-varying perturbations and near on-axis illumination.

We apply this formalism in chapter 5 first to the case of a circular pupil with central obscuration. We demonstrate reproduction of prior results before exploring the expanded behavior of the systems. We also demonstrate that this framework provides an explanation for the observed "bell-bagel" transition (Soummer et al., 2009). We then move to consideration of a non-circular pupil, a hypothetical coronagraph with the JWST pupil design. This irregular shape lets us demonstrate the flexibility and limitations of our approach.

Our conclusions are in chapter 6, along with speculation on possibilities to extend our work.

References

Wolszczan, A. and D. A. Frail (1992). "A planetary system around the millisecond pulsar PSR1257 + 12". In: *Nature* 355.6356, pp. 145–147.

Mayor, M. and D. Queloz (1995). "A Jupiter-mass companion to a solar-type star". In: *Nature* 378, pp. 355–359.

The Extrasolar Planets Encyclopaedia. URL: http://exoplanet.eu/.

Exoplanet Populations. URL: https://www.nasa.gov/sites/default/files/thumbnails/image/press-web25_exoplanet_populations.jpg.

"Exoplanets" (2014). In: *PNAS* 111.35. Ed. by Adam S. Burrowsa and Geoffrey W. Marcyb. DOI: 10.1073/pnas.1409934111.

Exoplanet Exploration: Historic Timeline. URL: https://exoplanets.nasa.gov/alien-worlds/historic-timeline/.

Charbonneau, David, Timothy M. Brown, David W. Latham, and Michel Mayor (2000). "Detection of Planetary Transits Across a Sun-like Star". In: *APJ Letters* 529.1.

Udalski, A., K. Zebrun, M. Szymanski, M. Kubiak, I. Soszynski, O. Szewczyk, L. Wyrzykowski, and G. Pietrzynski (2002). "The Optical Gravitational Lensing Experiment. Search for Planetary and Low-Luminosity Object Transits in the Galactic Disk. Results of 2001 Campaign - Supplement". In: *Acta Astron* 52.115. eprint: arXiv:astro-ph/0207133.

Kepler and K2. URL: https://www.nasa.gov/mission_pages/kepler/main/index.html.

Transiting Exoplanet Survey Satellite. URL: https://tess.gsfc.nasa.gov/.

TESS NasaFacts. URL: https://tess.gsfc.nasa.gov/documents/TESS_FactSheet_Oct2014.pdf.

James Webb Space Telescope. URL: https://jwst.nasa.gov/.

CHaracterising ExOPlanet Satellite. URL: http://sci.esa.int/cheops/.

The CHEOPS Mission. URL: http://cheops.unibe.ch/.

Traub, Wesley A. and Ben R. Oppenheimer (2010). "Direct Imaging of Exoplanets". In: *Exoplanets*. Ed. by S. Seager et. al.

2M1207b - First image of an exoplanet. URL: https://exoplanets.nasa.gov/resources/300/2m1207b-first-image-of-an-exoplanet/.

Spitzer Space Telescope. URL: http://www.spitzer.caltech.edu/.

NASA's Spitzer Marks Beginning of New Age of Planetary Science. URL: http://www.spitzer.caltech.edu/news/189-ssc2005-09-NASA-s-Spitzer-Marks-Beginning-of-New-Age-of-Planetary-Science.

Hubble Directly Observes Planet Orbiting Fomalhaut. URL: http://hubblesite.org/news_release/news/2008-39.

Wide-Field Infrared Survey Telescope. URL: https://wfirst.gsfc.nasa.gov/.

Planetary science wins big in NASA's new spending plan. URL: http://www.sciencemag.org/news/2018/03/planetary-science-wins-big-nasa-s-new-spending-plan.

WFIRST: Exoplanets - Direct Imaging. URL: https://wfirst.gsfc.nasa.gov/exoplanets_direct_imaging.html.

Exoplanet Probe to Medium Scale Direct Imaging Mission Requirements and Characteristics - (SAG9) Final report (2015). NASA SAG9 group. URL: https://exep.jpl.nasa.gov/files/exep/ExoPAG-SAG9-Final.pdf.

Lyot, B. (1939). "The Study of the Solar Corona and Prominences without Eclipses". In: *MNRAS* 99.8, pp. 580–594.

Guyon, O., E.A. Pluzhnik, M.J. Kuchner, B. Collins, and S.T. Ridgway (2006). "Theoretical Limits on Extrasolar Terrestrial Planet Detection with Coronagraphs". In: *ApJ Supplement Series* 167.1.

Spitzer, L. (1962). "The Beginnings and Future of Space Astronomy". In: *American Scientist* 50.3, p. 473.

Kuchner, Marc J. and Wesley A. Traub (2002). "A Coronagraph with a Band-limited Mask for Finding Terrestrial Planets". In: *ApJ* 570.2.

Roddier, F. and C. Roddier (1997). "STELLAR CORONOGRAPH WITH PHASE MASK". In: *Publications of the Astronomical Society of the Pacific* 109.1737.

Rouan, D., P. Riaud, A. Boccaletti, Y. Clénet, and A. Labeyrie (2000). "The Four-Quadrant Phase-Mask Coronagraph. I. Principle". In: *Publications of the Astronomical Society of the Pacific* 112.777.

Foo, Gregory, David M. Palacios, and Grover A. Swartzlander (2005). "Optical vortex coronagraph". In: *Optics Letters* 30.24.

Jacquinot, P. and B. Roizen-Dossier (1964). "Progress in Optics". In: vol. 3. Progress in Optics. North-Holland Publishing Company. Chap. 2. Apodization.

Kasdin, N. Jeremy, Robert J. Vanderbei, David N. Spergel, and Michael G. Littman (2003). "Extrasolar Planet Finding via Optimal Apodized-Pupil and Shaped-Pupil Coronagraphs". In: *ApJ* 582.2.

Aime, Claude, Rémi Soummer, and A Ferrari (2002). "Total coronagraphic extinction of rectangular apertures using linear prolate apodizations". In: 389.

Soummer, Rémi (2005). "Apodized Pupil Lyot Coronagraphs for Arbitrary Telescope Apertures". In: *ApJ Letters* 618.2.

Guyon, Olivier, Eugene A. Pluzhnik, Raphael Galicher, Frantz Martinache, Stephen T. Ridgway, and Robert A. Woodruff (2005). "Exoplanet Imaging with a Phase-induced Amplitude Apodization Coronagraph. I. Principle". In: *ApJ* 622.1.

Soummer, R., L. Pueyo, A. Ferrari, C.Aime, and A. Sivaramakrishnan (2009). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures, II. Theoretical Properties and Application to Extremely Large Telescopes". In: *The Astrophysical Journal* 695.1, pp. 695–706.

Chapter 2

Mathematics of the Slepian problem

Before we approach the coronagraph, we will develop the mathematics and formalism used. While this will require a small amount of explanation of the behavior of a coronagraph, we defer the more detailed version to § 3.1.

It was first noted by (Aime, Soummer, and Ferrari, 2002) that apodization of the rectangular pupil for solid (Lyot-style) and phase-inducing (Roddier and Roddier, 1997) rectangular masks in a coronagraph was a mathematical problem that had previously been solved. That it applied to circular pupils and masks, and its results there, was shown in (Soummer, 2005). That paper also noted that these were specific cases of the mathematical problem which had been solved in principle.

The reasoning that lead to this discovery was as follows: the relevant portion of the coronagraph is arranged in a series of Fourier conjugate planes. The first is the pupil itself; the next contains the mask; and the third will carry the Lyot stop. For the purposes of the attack, we do not need to consider

the shape of that stop. With such a set-up, an initial field ϕ from the pupil plane will become the convolution $\phi \star (1 - \epsilon \widehat{M})$ in the third (Lyot) plane. The symbol ϵ is a factor of one or two, due to the presence of a solid or phase mask respectively, while $\widehat{M}$ is the Fourier transform of the mask. If the convolution $\phi \star \widehat{M} = \Lambda\phi$, then the intensity in this plane is proportional to $(1 - \epsilon\Lambda)^2$ and near-total suppression can occur.

Mathematically speaking, this means that we desire a function which is an eigenfunction of repeated Fourier transforms, each with a limitation in the domain of integration. We illustrate this schematically in figure 2.1.

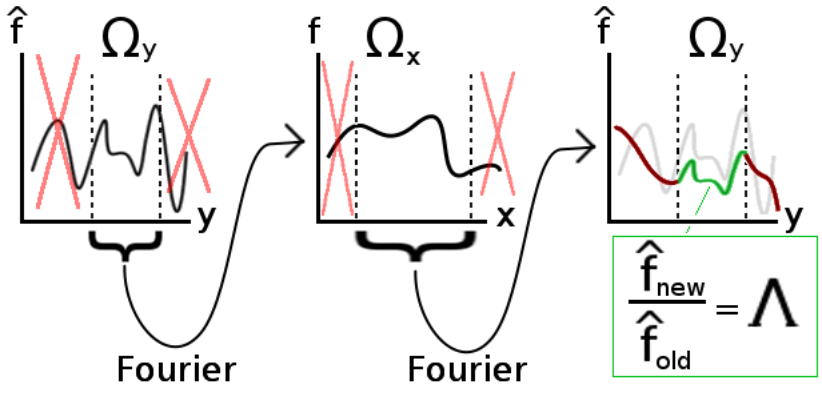


Figure 2.1: Illustration of the "Slepian" problem: repeated truncation and Fourier transforms which produce a scaled version of the original function within some domain of interest.

This problem was first systematically studied by Dr. David Slepian et. al. ((Slepian and Pollak, 1961), (Landau and Pollak, 1961), (Landau and Pollak, 1962), (Slepian, 1964), (Slepian, 1978), (Slepian, 1976)). His original focus was on signals which were recorded in limited time and frequency, especially

how smoothing windows redistributed power in the frequency bands. This mathematics has since been generalized to higher dimensions (Slepian, 1964) and other signal types (e.g., (Simmons and Dahlen, 2006)).

In this chapter, we will review and restate the mathematics of this problem, in a notation designed to focus on the abstract linear-algebra properties of the functions in question. § 2.1 traces Slepian's logic to pose the problem in a well-defined mathematical manner. Section 2.2 introduces the new formalism used in the rest of this thesis, which is taken from the Dirac "bra-ket" notation for quantum mechanics. § 2.3 then restates the Slepian problem in a manner which emphasizes the nature of the relevant linear operator, called the *kernel*.

§ 2.4 shows that this operator belongs to a general type called *trace-class* operators. This allows us to state a large number of general properties, as these objects have been well-studied before (e.g. (Birman and Solomjak, 1987)). Section 2.5 demonstrates that the kernel is, in a sense, dual to another kernel, and develops the relation between them. This duality is the key to our ability to apply these mathematics to the coronagraphic problem in a novel fashion.

§2.6 lists the desirable properties for a "working" basis when attacking any such problem. These are the functions which serve as the building blocks which are actually manipulated to solve a problem at hand. We conclude in 2.7.

2.1 Initial Setup

Schwartz's work (Schwartz, 1952) on the Paley-Wiener theorem of harmonic analysis (Paley and Winer, 1934) for smooth functions means that no function

or distribution can be simultaneously constrained to a compact[1] region in its domain and in the domain of its Fourier transform (Stein and Weiss, 2016). Thus, any function which is only defined in some finite window in position space has some spread in wave-number space, and vice-versa. This is the generic mathematical statement which lies behind the familiar Heisenberg Uncertainty Principle (Papoulis, 1968).

Since all signals are in some way truncated in the domain of measurement, their measured Fourier transform is the convolution of the true Fourier transform of the underlying signal and that of the truncating window. Since the latter is always constrained to a compact region, the measured transform of the signal extends to all portions of the transform domain. The result is that the Fourier transform of the recorded signal is inevitably distorted.

A weighted windowing function offers the possibility of altering this bleed-through. An obvious goal is then to find a weighting which is well-concentrated in some small domain of Fourier space, while still preserving the complete limitation in the original space. This was the motivation for Slepian (Slepian and Pollak, 1961), who was concerned with minimizing the effects of limited-time recording of signals on its transform to frequency. We here follow his process from (Slepian and Pollak, 1961), (Landau and Pollak, 1961), and (Slepian, 1964), though in slightly different notation.

We start with some signal, a function $f(\mathbf{x})$, where $\mathbf{x}$ is an $n-$dimensional real vector. (The most obvious examples are position, with $n = 3$ or time, $n = 1$.) We assume that we are interested in the behavior of f on a compact

[1] A *compact* region in $n-$dimensional Euclidean space is one that is *closed* (containing all limit points) and *bounded* (of finite maximum distance between any two points in the region).

region Ω_x; the same symbol $\Omega_x(\mathbf{x})$ will also be used as the indicator function, taking the value 1 inside the region, and zero outside.

The Fourier transform of the function is defined as

$$\widehat{f}\,(\mathbf{y}) \equiv \int \mathrm{d}^n x\, e^{-i\mathbf{y}\cdot\mathbf{x}} f\,(\mathbf{x})$$

where $\mathbf{y}$ is the Fourier-space vector. (If $\mathbf{x}$ is position, $\mathbf{y}$ is wavenumber; if x is time, y is circular frequency, etc.) The inverse Fourier transform is given by

$$f\,(\mathbf{x}) = \int \mathrm{đ}^n y\, e^{i\mathbf{x}\cdot\mathbf{y}} \widehat{f}\,(\mathbf{y})$$

where for cleanliness the notation $\mathrm{đ}^n y \equiv \frac{1}{(2\pi)^n}\,\mathrm{d}^n y$ is used to hide factors of 2π. Whatever units $\mathbf{x}$ may have, $\mathbf{y}$ carries their inverse so that the product $\mathbf{x}\cdot\mathbf{y}$ is a pure number.

We will refer to $|f|^2$ as a density, which produces a total quantity contained in region Ω_x of $\int \mathrm{d}^n x\, \Omega_x(\mathbf{x}) |f(\mathbf{x})|^2$. (Electric fields producing an energy density, or a quantum wavefunction producing a probability density, are common examples.) This density is usually referred to as a "energy" density for convenience's sake.

The Fourier domain of interest is denoted by Ω_y, which we again take to be compact. Again, it is a basic result of Fourier analysis that there exists no function that is completely contained in both Ω_x and Ω_y simultaneously. Instead, we study the relative concentration of the function. Our quantity of interest is the amount of energy inside Ω_y, after limiting the function to Ω_x, compared to the total energy without any limitations.

The energy in the function without limits is given by the familiar $\int \mathrm{d}^n x\, |f(\mathbf{x})|^2$,

which is also equal to $\int đ^n y\, |\widehat{f}(\mathbf{y})|^2$ by Parseval's theorem. If we wanted to look at the energy U in Ω_x without concentration in Ω_y, then we have

$$\begin{aligned} U &= \int \mathrm{d}^n x\, \Omega_x(\mathbf{x}) |f(\mathbf{x})|^2 \\ &= \int \mathrm{d}^n x\, \Omega_x(\mathbf{x}) \int đ^n y\, \widehat{f}(\mathbf{y}) e^{i\mathbf{y}\cdot\mathbf{x}} \int đ^n y'\, \widehat{f}^*(\mathbf{y}) e^{-i\mathbf{y}'\cdot\mathbf{x}} \end{aligned}$$

by the definition of the inverse Fourier transform. ($\widehat{f}^*$ denotes the complex conjugate.)

The form above shows us how to generalize to measuring the energy left in Ω_x after the Fourier-transform is limited to Ω_y:

$$\int \mathrm{d}^n x\, \Omega_x(\mathbf{x}) \int đ^n y\, \widehat{f}(\mathbf{y}_1) \Omega_y(\mathbf{y}_1) e^{i\mathbf{y}_1\cdot\mathbf{x}} \int đ^n y_2\, \Omega_y(\mathbf{y}_2) \widehat{f}^*(\mathbf{y}_2) e^{-i\mathbf{y}_2\cdot\mathbf{x}}$$

We want the ratio of this limited-domain energy to the total energy, known as the *Rayleigh quotient*:

$$\frac{\int đ^n y_1 \int đ^n y_2\ \widehat{f}(\mathbf{y}_1) \left(\Omega_y(\mathbf{y}_1) \int \mathrm{d}^n x\, \Omega_x(\mathbf{x})\ e^{i\mathbf{x}\cdot(\mathbf{y}_1-\mathbf{y}_2)} \Omega_y(\mathbf{y}_2) \right) \widehat{f}(\mathbf{y}_2)}{\int \mathrm{d}^n x\, |f(\mathbf{x})|^2} \tag{2.1}$$

The quantity in parentheses is called the *kernel*, written as $K(\mathbf{y}_1, \mathbf{y}_2)$. We can consider it to be a measure of correlation between the points $\mathbf{y}_1$ and $\mathbf{y}_2$ induced by the $x-$space truncation. If Ω_x were the identity, not restricting x, then the kernel would be $\Omega_y(\mathbf{y}_1)\delta(\mathbf{y}_2 - \mathbf{y}_1)\Omega_y(\mathbf{y}_2)$. No correlation would exist, and all points would have equal weighting.

The kernel is a linear operator, though it has infinite indices $(\mathbf{y}_1, \mathbf{y}_2)$ where a regular matrix has countable indices (i, j). The linear-algebra properties of kernels are studied in more detail in 2.4. For now, we focus on the fact that *if*

there are eigenfunctions $\widehat{\phi}_a(\mathbf{y}_2)$ with eigenvalues Λ_a, such that

$$\int \mathrm{d}^n y_2 \, K(\mathbf{y}_1, \mathbf{y}_2) \widehat{\phi}_a(\mathbf{y}_2) = \Lambda_a \widehat{\phi}_a(\mathbf{y}_1)$$

then the Rayleigh quotient of such an eigenfunction is just

$$\Lambda_a \frac{\int \mathrm{d}^n y \, |\widehat{\phi}_a(\mathbf{y})|^2}{\int \mathrm{d}^n x \, |\phi_a(\mathbf{x})|^2} = \Lambda_a$$

by using Parseval's theorem. In chapter 3, we will show how just such a kernel is used for describing the pupil-mask portion of the coronagraph.

These eigenfunctions, which we will also refer to as "Slepian" functions, are only scaled under the combined operation of limiting to Ω_y, Fourier transforming, limiting to Ω_x, Fourier transforming, and considering only those regions in Ω_y after this second transform. The eigenvalues measure the degree of concentration in Ω_x, and the largest Λ_a corresponds to the "best-concentrated" Slepian function $\phi_a(\mathbf{x})$. We will always assume that we have ordered the eigenvalues, and their corresponding eigenfunctions, in descending order.

2.2 Change of notation

All of the above discussion was written in notation similar to that used by (Slepian and Pollak, 1961), which emphasizes the *integral* nature of the operations involved. This allowed finding explicit solutions to the problem in simple cases of interest. For a one-dimensional connected window the eigenfunctions are the "angular prolate spheroidal wave functions" (Slepian and Pollak, 1961). In a two-dimensional annular region, the eigenfunctions are the

"generalized prolate spheroidal functions" (Slepian, 1964).

We have found it useful to change our notation to instead emphasize the *linear* nature of the operations. Table 2.1 summarizes the new notation. This has several advantages: it is more compact; it encourages thought on the abstract mathematical properties of the system; and it prepared us for the transition to finite-matrix approximation. We will introduce this notation by reviewing some basics of functional analysis for the functions we will use in this thesis. All of the mathematical statements in this section may be found in an introductory text on that topic.

Functional form	Abstract form
$f(\mathbf{x})$	$\|f\rangle$
$\widehat{f}(\mathbf{y}) \equiv \mathfrak{F}[f(\mathbf{x})](\mathbf{y})$	$\|f\rangle$
$\Omega_y(\mathbf{y}_1, \mathbf{y}_2)$	P_2
$\Omega_x(\mathbf{x}_1, \mathbf{x}_2)$	P_1
$K(\mathbf{y}_1, \mathbf{y}_2) \equiv \Omega_{y_1} \left[\int \text{đ}^n x \, \Omega_x e^{i\mathbf{x}\cdot(\mathbf{y}_1 - \mathbf{y}_2)} \right] \Omega_{y_2}$	$P_2 P_1 P_2$
$\int \text{đ}^n y \, [f(\mathbf{y})]^* [g(\mathbf{y})]$	$\langle f \| g \rangle$
$\int \text{đ}^n y_2 \, K(\mathbf{y}_1, \mathbf{y}_2) \widehat{\phi}_a(\mathbf{y}_2) = \Lambda_a \widehat{\phi}_a(\mathbf{y}_1)$	$K \|a\rangle = \Lambda_a \|a\rangle$

Table 2.1: Dictionary for new notation. Summation is implied on repeated indices.

It is a starting point for functional analysis that the additive nature of functions, $c\,[f(\mathbf{x}) + g(\mathbf{x})] = [c \times f(\mathbf{x})] + [c \times g(\mathbf{x})]$ means that they can potentially be treated as elements of an abstract vector space. We will consider only "smooth" functions: any $n-$th derivative is finite, with only some countable number of discontinuities. This is a common set of functions to consider for physical settings. It also includes such objects as the Dirac delta "function,"

more properly called a distribution.

Descriptions of a function in $\mathbf{x}$ or in the Fourier space $\mathbf{y}$, while different in form, still refer to the same abstract function. This is the same reasoning that switching coordinates to a new, rotated basis only provides a new description of the same vector. While the explicit values of the components change, we retain a symbol such as $\mathbf{v}$ to represent the vector regardless of what basis is used.

In similar form, we will use Dirac's bracket notation $|f\rangle$ to represent a function, without regard to any basis in which it can be expressed. There exist the conjugate transposes $\langle f|$ of vectors. If we are dealing with complex-valued functions, then this truly will be a complex conjugate. The notation allows us to write the inner (dot) product of two vectors as $\langle f|f\rangle$, when such an inner product exists. We will only be considering functions for which this is true, and in some particular representation $f(\mathbf{x})$ is written as $\int \mathrm{d}^n x\, f^*(\mathbf{x})\, f(\mathbf{x})$.

If we further restrict our study to functions where the inner product be finite, we have reached a familiar vector space. Hilbert and others developed the study of this space of functions, which is known as L^2.[2] Since so many different applications require these functions, it has been studied extensively (*L2 Space*).

To summarize the restrictions on our functions: they must be finite, infinitely differentiable (up to a countable number of discontinuities, as occurs in a Heaviside step function), and have finite inner product with any element

[2] There are more Hilbert spaces than just L^2. Their requirements are that they are a vector space with an inner product. The inner product must satisfy a few basic laws, and the vector space must be complete in the Cauchy sense. Hilbert spaces are the basic example for functional analysis.

of the space. Addition of functions, or multiplication by a complex number, produces a function that is still within our consideration.

We have focused here on L^2 as the electric fields in our coronagraph are very well described as being smooth, and any apodization production will itself produce a smooth function. Additional restrictions on the space of functions that we will use for the apodization are discussed in § 3.3. The rest of this chapter will be general to L^2.

A complete basis of vectors is one which *spans* the vector space. That is, if $\{|b_i\rangle\}$ is such a set, then any function $|f\rangle$ from the space can be decomposed in this basis as $\sum_i |b_i\rangle \langle b_i|f\rangle$. This is no different from any familiar vector being $\mathbf{v} = (\mathbf{v} \cdot \hat{x})\hat{x} + (\mathbf{v} \cdot \hat{y})\hat{y}$.

The basis vectors can be labeled by a continuous index. If we are using some coordinates $\mathbf{r}$, then the value of the function at any point $\mathbf{r}^*$ written as $f(\mathbf{r}^*) = \langle \mathbf{r}^*|f\rangle$, and the function itself represented as $|\mathbf{r}\rangle \langle \mathbf{r}|f\rangle$, summed over all $\mathbf{r}$.

We must be careful when writing the product of functions into this notation. We would be greatly helped if there exists a rule for turning the product of two basis functions into a weighted sum of basis functions, $|i\rangle \otimes |j\rangle = \sum_k c_k^{ij} |k\rangle$. This structure converts the vector space into an abstract *algebra*. The specific weighting coefficients would be determined by additional restraints on the space of functions. This is somewhat useful for products in image space, on the mask, but not in pupil space. See section 3.3.3 for further details.

An operator, $\mathbb{O}$, is the equivalent of a matrix. They act on vectors as $\mathbb{O}|f\rangle$, producing a new vector. They also may act on other operators $\mathbb{O}_1\mathbb{O}_2$,

producing a new composite operator. Their representation requires two basis vectors, $O_{ij} = \langle b_i | \mathbb{O} | b_j \rangle$ from the chosen basis. As with vectors, we refer to the O_{ij} as the components of the operator in this basis.

One special operator is the identity operator. It is written as

$$\mathbb{I} = |b_i\rangle \langle b_i| \tag{2.2}$$

with an implicit sum over all $|b_i\rangle$. Al components are one.

Another operator is the Fourier transform $\mathfrak{F} |f\rangle$, which represents a change of basis:

$$\begin{aligned} \mathfrak{F} |x\rangle \langle x|f\rangle &= |y\rangle \langle y|x\rangle \langle x|f\rangle \\ &= |y\rangle \langle y|f\rangle \end{aligned} \tag{2.3}$$

assuming that x and y are the labels for the Fourier-conjugate coordinates. $\langle y|x\rangle$ is the familiar $e^{-i\mathbf{x}\cdot\mathbf{y}}$, and the sum is shorthand for $\int \mathrm{d}^n y$. The transform is therefore in this abstract sense just the identity operator.

2.3 New notation with the Slepian problem

Let us now return to Slepian's problem using this new notation. In the Rayleigh quotient (2.1), the denominator is just the inner product of f with itself:

$$\int \mathrm{d}^n x \, f^* (\mathbf{x}) \, f (\mathbf{x}) = \sum_x \langle f|x\rangle \langle x|f\rangle = \langle f|f\rangle$$

In the numerator, we must analyze the situation a bit before implementing the new notation. $\Omega_x (\mathbf{x})$ and $\Omega_y (\mathbf{y})$ are projection operators P_1 and P_2 which

limit us to subspaces Ω_x and Ω_y. They have their natural expression in the x or y basis, but do not have to be written in such.

Formally, since they are operators, they should be written as $\Omega_x(\mathbf{x}_1, \mathbf{x}_2)$, being $\delta(\mathbf{x}_1 - \mathbf{x}_2)$ if both $\mathbf{x}$ values are in the region Ω_x and zero otherwise. The extra coordinate is integrated over, since this represents the inner product. This substitution amounts to $\Omega_x(\mathbf{x}) \to \int \mathrm{d}^n x' \, \Omega_x(\mathbf{x}, \mathbf{x}')$, and likewise for the Ω_y. We will usually neglect this level of detail other than in this derivation.

The whole numerator is therefore

$$\int đ^n y_1 đ^n y_1' \;\; \mathrm{d}^n x \;\; \mathrm{d}^n x' đ^n y_2' đ^n y_2$$

$$\left[\widehat{f}^*(\mathbf{y}_1)\Omega_y(\mathbf{y}_1, \mathbf{y}_1')e^{-i\mathbf{y}_1'\cdot\mathbf{x}}\Omega_x(\mathbf{x}, \mathbf{x}')e^{i\mathbf{x}'\cdot\mathbf{y}_2'}\Omega_y(\mathbf{y}_2', \mathbf{y}_2)\widehat{f}(\mathbf{y}_2)\right]$$

We may now rewrite this in the abstract notation. The integrals represent the sums over the relevant coordinates. $\widehat{f}(\mathbf{y}_2) = |y_2\rangle \langle y_2|f\rangle$, while $\widehat{f}^*(\mathbf{y}_1) = \langle f|y_1\rangle \langle y_1|$ to represent the complex conjugation. The quantity $e^{i\mathbf{y}_1'\cdot\mathbf{x}}$ is just the change of basis element $\langle y_1'|x\rangle$. Putting these together, we have

$$\langle f|y_1\rangle \langle y_1| P_2 \left|y_1'\right\rangle \left\langle y_1'|x\right\rangle \langle x| P_1 \left|x'\right\rangle \left\langle x'|y_2'\right\rangle \left\langle y_2'\right| P_2 |y_2\rangle \langle y_2|f\rangle$$

with sums over all paired variables implied.

If use our relation for the Fourier transform (2.3), and remember that it is just equal to the identity, then all of these paired variables just become one and can be removed. Reducing this leaves us with the much simpler expression for (2.1),

$$\frac{\langle f| P_2 P_1 P_2 |f\rangle}{\langle f|f\rangle} \tag{2.4}$$

We can now see that our original kernel operator $K(\mathbf{y}_1, \mathbf{y}_2)$ is the expression of this triple projection operator

$$K = P_2 P_1 P_2 \tag{2.5}$$

in the bases naturally suited for it. We are finding the eigenvalues and eigenfunctions (eigenvectors) $K|a\rangle = \Lambda_a |a\rangle$ of the kernel. We will sometimes refer to these eigenfunctions as "Slepian" functions.

While we independently derived this triple-projection identity, later research showed that it has previously been partially recognized in (Landau and Pollak, 1961). We have only found it used in (Simmons and Dahlen, 2006), although the one- and three-dimensional Slepian functions for line intervals are commonly used in signal analysis. So far as we are aware it is unknown in coronagraphic literature.

2.4 Abstract kernel properties

Before we look at any coronagraphic kernel, we will examine the properties of all Slepian style kernels.

It is obvious that the kernel is Hermitian: $K = (K^{\mathrm{T}})^* \equiv K^\dagger$. It is the product of projection operators, which are bounded[3] positive[4] operators; therefore, K is a bounded positive operator. These properties together imply (Hogan and Lakey, 2012) that the kernel is a "nuclear" operator. Since it is defined in a Hilbert space, it has a well-defined trace. It is therefore a "trace-class"

[3] finite eigenvalues
[4] positive eigenvalues, neglecting the null space

operator (Gosson, 2011). Our following facts for such operators are taken from (Birman and Solomjak, 1987) (unless otherwise noted).

The eigenvalues and eigenvectors exist and can be labeled with an integer a in descending order of eigenvalue: $\Lambda_1 \geq \Lambda_2 \geq \Lambda_3 \geq \ldots$, with corresponding eigenvectors $|1\rangle, |2\rangle, |3\rangle, \ldots |a\rangle, \ldots$. The eigenfunctions corresponding to different eigenvalues are orthogonal, $\langle a|b\rangle \propto \delta_{ab}$. We will normalize our eigenvectors so that $\langle a|a\rangle = 1$. Eigenfunctions for identical eigenvalues can be arbitrarily diagonalized using the Gram-Schmidt procedure. Every eigenfunction is wholly contained within P_2, so that $P_2 |a\rangle = |a\rangle$ and $(\mathcal{I} - P_2) |a\rangle = 0$.

Denoting the matrix 2−norm $||\cdot||$, $||P_1|| = ||P_2|| = 1$ and so our kernel's norm $||K|| \leq 1$. [5] For $||K||$ to be one, there must be some $|q\rangle$ such that $|K |q\rangle| = ||q\rangle|$. This would require both $|P_1 |q\rangle| = ||q\rangle|$ and $|P_2 |q\rangle| = ||q\rangle|$, so that $P_1 P_2 |q\rangle = P_2 P_1 |q\rangle$. P_1 and P_2 would thus weakly commute, which is the trivial case we do not want to study.

Thus, $||K|| < 1$. Since the operator norm bounds the magnitudes of the eigenvalues (*Matrix Norm*) (Birman and Solomjak, 1987), the maximal eigenvalue $\Lambda_1 < 1$. Because K is a bounded positive operator, the smallest eigenvalue must be greater than zero. (This does not rule out an infinitesimal eigenvalue.) We safely have that $1 > \Lambda_1 \geq \Lambda_2 \geq \ldots > 0$ for all Slepian-style kernels.

For nuclear operators, $\sum \Lambda_a < \infty$ is guaranteed, as well as $|\sum \Lambda_a^p|^{1/p}$ for all powers $p \geq 1$ (*An Elementary Proof of the Spectral Radius Formula for Matrices*).

[5] The matrix 2−norm is defined as $||M|| = \max_{\mathbf{v}} |M\mathbf{v}|/|\mathbf{v}|$. All matrix norms obey $||M_1 M_2|| \leq ||M_1|| \, ||M_2||$. (*Matrix Norm*).

Our dependence on a must therefore decay faster than any polynomial; e.g., exponential e^{-a} or factorial $1/a!$. An infinite number of eigenvalues cluster arbitrarily close to zero. These are those eigenfunctions almost wholly outside the regions of P_1 and P_2.

A theorem due to Lidskii (Birman and Solomjak, 1987) states that for trace-class operators, the "matrix" trace (sum of eigenvalues) is invariant of representation. In particular, it is equal to the spectral trace,

$$\begin{aligned} \mathrm{Tr}_{spec} &= \int \text{đ}^n y \, \Omega_y K(y,y) \\ &= \int \text{đ}^n y \, \Omega_y \int \mathrm{d}^n x \, \Omega_x \\ &= \frac{|\Omega_x||\Omega_y|}{(2\pi)^n} \end{aligned}$$

where $|\Omega_i|$ is the area or volume of the domain in question. Therefore:

$$\sum_{i=1}^{\infty} \Lambda_i = \frac{|\Omega_x||\Omega_y|}{(2\pi)^n} \tag{2.6}$$

for *all* Slepian-style problems, regardless of the shapes of the domains or the dimensionality of the problem. This is a generalization of the $2WT$ theorem from signal analysis (Slepian and Pollak, 1961), also realized in (Simmons and Dahlen, 2006).

The trace is an approximation to the number of eigenvectors of $\Lambda_a \approx 1$, which are the eigenvectors mostly inside the region of interest and therefore expected to be of must use to us. Describing functions as a sum of more than this number of basis functions or eigenfunctions is mostly redundant. In this context, the trace is sometimes called the "Shannon number" (Simmons and

Dahlen, 2006). As a practical measure, this provides an immediate error check for any generated kernel.

The kernel itself can be written as $\sum_a \Lambda_a |a\rangle \langle a|$, and $P_2 = \sum_a |a\rangle \langle a|$. This means that the eigenfunctions span the space, as described in 2.2. Any function inside of Ω_y can be written as a sum of the eigenfunctions. Functions of K can be defined using Taylor expansions, and any power $K^\nu = \sum_a |a\rangle \langle a| \Lambda_a^\nu$.

Finite dimensional representations of our kernel, $\sum_{a=1}^{N} \Lambda_a |a\rangle \langle a|$, are "dense" in the sense that we can find some N for which $\mathrm{Tr}\,(K - K_N) \approx \Lambda_{N+1} \leq \varepsilon$ for any desired ε. This means that the "energy" lost from not considering these higher functions above N can be made as small as desired. We can therefore describe the infinite-dimensional kernel with a finite matrix without introducing more than a small amount of error, of order Λ_{N+1}.

2.5 Dual kernel

Our kernel was $K \equiv P_2 P_1 P_2$. We can consider the natural "dual" kernel to be $K' \equiv P_1 P_2 P_1$. This duality will prove to be key to our approach to the coronagraph. While this consideration is novel so far as we know, the conclusions are not so, being anticipated as far back as Slepian's original work.

Recall that K describes functions entirely in Ω_y (unaffected by P_2) and well-concentrated in Ω_x; K' therefore describes functions completely in Ω_x (unaffected by P_1) and well-concentrated in Ω_y. The dual's eigensystem is written as $|a'\rangle$ and Λ'_a.

Let us decompose the original vector space using a basis which splits into two parts. Inside Ω_x, we have $\{|i_A\rangle\}$, using A to indicate that the basis

vector $|i\rangle$ belongs to this part of the split. Outside of Ω_x, the basis is $\{|i_B\rangle\}$, using B similarly. This definition is equivalent to saying that $P_1 |i_A\rangle = |i_A\rangle$ and $P_1 |i_B\rangle = 0$ for all $|i_A\rangle$ and $|i_B\rangle$. Since these two collections completely partition all of the space, we say that the whole space is the direct sum $\{|i_A\rangle\} \oplus \{|i_B\rangle\}$.

We want to see how the kernels appear when written in this division of space. To do so, we will look at the difference when we apply KK' compared to $K'K$. This difference, $KK' - K'K$, is known as the commutator of the operators and written as $[K, K']$. It is itself an operator. Putting in the definitions for K and K', $[K, K'] = (P_2P_1)^3 - (P_1P_2)^3$. The matrix elements for the commutator are then

$$\begin{aligned} \langle i_{A1}| [K, K'] |i_{A2}\rangle &= 0 \\ \langle i_A| [K, K'] |i_B\rangle &= -\langle i_A| P_2K'P_2 |i_B\rangle \\ \langle i_{B1}| [K, K'] |i_{B2}\rangle &= 0 \end{aligned}$$

We then examine the same matrix elements, but this time use a basis designed for splitting the vector space into pieces inside and outside of P_2 instead of P_1. We will use the lettering $|j\rangle$, with $|j_A\rangle$ inside of P_2 and $|j_B\rangle$ outside. In this division, only

$$\langle j_A| [K, K'] |j_B\rangle = \langle j_A| P_1KP_1 |j_B\rangle$$

is not zero.

In both of these cases, we found that the elements of the commutator $[K, K']$ are always zero when we restrict ourselves both vectors only from $\{|i_A\rangle\}$, or

both from $\{|i_B\rangle\}$, or likewise from $\{|j_A\rangle\}$ or $\{|j_B\rangle\}$. A block representation of this is (2.7). [6]

$$\left[\begin{pmatrix} K_{A,A} & K_{A,B} \\ K_{B,A} & K_{B,B} \end{pmatrix}, \begin{pmatrix} K'_{A,A} & K'_{A,B} \\ K'_{B,A} & K'_{B,B} \end{pmatrix}\right] = \begin{pmatrix} 0 & \text{NOT ZERO} \\ \text{NOT ZERO} & 0 \end{pmatrix} \tag{2.7}$$

When two operators commute, $[K, K'] = 0$, then it is a basic fact of linear algebra that they have the same eigenvectors up to an overall constant factor, though the eigenvalues will be different. In our case, this means that the restriction of the eigenvectors to P_1 and P_2 are identical. If we look at the eigenvectors of K outside of P_1, however, then they are *not* the same as those of K'. Figure 2.2 gives a schematic example.

$$|a\rangle = \begin{pmatrix} |a\rangle_A \\ |a\rangle_B \end{pmatrix} \qquad |a\rangle_A \propto |a'\rangle_A$$

$$|a'\rangle = \begin{pmatrix} |a'\rangle_A \\ |a'\rangle_B \end{pmatrix} \qquad |a\rangle_B \not\propto |a'\rangle_B$$

Since this is the case, we will start using $|a\rangle$ to refer to the part of the eigenvectors inside the spaces of interest, which we had been calling $|a\rangle_A$. Unless otherwise stated, this is the case for the rest of this thesis.

Let's look at the proportional eigenvectors, writing $P_2 |a\rangle = cP_2 |a'\rangle$ to

[6] The formal language is that in the relevant subspaces, the two K and K' are "weakly commutative operators."

mean that the ath eigenvectors of K and K' are proportional.

$$\begin{aligned}
\langle a| P_2 P_2 |a\rangle &= c^2 \langle a'| P_2 P_2 |a'\rangle \\
\langle a|a\rangle &= c^2 \langle a'| P_1 P_2 P_1 |a'\rangle \\
1 &= c^2 \Lambda_{a'}
\end{aligned}$$

Therefore $|a\rangle = (\Lambda_{a'})^{-1/2} P_2 |a'\rangle$. The same reasoning gives us that $|a'\rangle = (\Lambda_a)^{-1/2} P_1 |a\rangle$.

$$\begin{aligned}
\frac{P_2 |a'\rangle}{\sqrt{\Lambda'_a}} &= |a\rangle \\
\frac{P_1 |a\rangle}{\sqrt{\Lambda_a}} &= |a'\rangle
\end{aligned} \tag{2.8}$$

Combining these two facts, the *ath eigenvalues of the kernel and its dual are identical.*

$$\Lambda_{a'} = \Lambda_a \tag{2.9}$$

We also have that $|a\rangle$ and $|a'\rangle$ can seem to extend the other outside their original P_2 or P_1, matching all boundary values. This seeming extension maintains the orthogonal nature of the eigenvectors. The values that this extension gives are *not* those that the actual eigenvector of the full kernel or its dual would take, if an analytical extension of their functional form exists; again, this is shown in figure 2.2.

These relations also mean that $P_1 = \sum |a'\rangle \langle a'| = \sum \frac{P_1 |a\rangle \langle a| P_1}{\Lambda_a}$. While this appears recursive, it will prove useful later on in 4.3. Interestingly, this has the form $P_1 K^{-1} P_1$, though we do not know of any major applications of this fact.

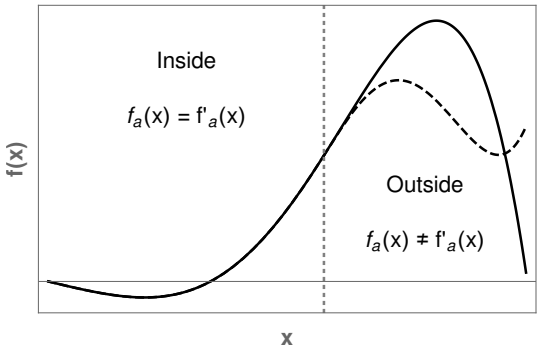


Figure 2.2: Illustration of eigenfunctions identical within a region, but the extension of one does not match the extension of the other outside that region.

The major use for these facts is that *if Ω_x is complicated, but the other region Ω_y is simple, we can consider the dual kernel instead. The eigenvalues are the same, and the eigenvectors in the areas of interest are proportional via* (2.8).

2.6 Basis considerations

The abstract nature of this statement of the problem points us to consider what the most convenient basis $\{b_j(\mathbf{x})\} = \{\sum_x |x\rangle \langle x|j\rangle\}$ is to express our Slepian functions. Each of the projection operators P_1 and P_2 has its own natural basis, so a good choice of $\{|j\rangle\}$ will in some sense align closely with one of these natural bases. Our ideal is a basis for which $P_2 |j\rangle = |j\rangle, (1 - P_2) |j\rangle = 0$, as we are interested in eigenvectors satisfying the same relation for $|a\rangle$. We will assume that $\langle j|k\rangle = \delta_{jk}$.

The basis' expression in $x-$ and $y-$coordinates should have nice analytical properties. Otherwise, calculations can become impractical, and insight is difficult.

We also desire a countable basis, so that we can approximate the kernel as a matrix and drastically simplify our problem. Such a countable basis should have a natural means of truncation, so that we may choose our size for the $N-$component approximation of the problem.

We have no guarantee that there exists a basis with all of these desired properties.

In whatever basis we choose, the kernel elements will be calculated by

$$K_{jk} \equiv \langle j| K |k\rangle = \int \mathrm{d}^n x \, \Omega_x (\mathbf{x}) \, b_j^* (\mathbf{x}) \, b_k (\mathbf{x}) \tag{2.10}$$

The eigenfunctions are then written as

$$\phi_a (\mathbf{x}) = \sum_j V_{a,j} \, b_j (\mathbf{x}) \tag{2.11}$$

where $V_{a,j} \equiv \langle j|a\rangle$ is the vector descriptions of the eigenfunction in that basis, found by solving the eigenproblem for the finite approximation K_{jk}.

There are three sources of error, $\max \left| \phi_a(\mathbf{x}) - \phi_a^{(N)} (\mathbf{x}) \right|$, introduced from our approximations: the finite truncation of the basis, the integral calculation of the kernel elements, and the choice of eigensystem algorithm. The latter two have known behavior from standard numerical routines, so the error from the first must be understood. This is our "natural means of truncation," and will vary from problem to problem with our choice of b_j.

2.7 Summary

We have introduced abstract notation for the analysis of our problem, as summarized in table 2.1. This notation allowed us to find a simple expression for the key operator, the *kernel* 2.5. The properties of the kernel allowed us to recognize it as belonging to a type known as "trace-class" operators, which are well-studied. In particular, we know that they have well-defined eigenvalues limited to between zero and one, with orthonormal eigenfunctions.

Ordering the eigenvalues, we have that they must decay nearly exponentially, and so only a few will be useful for us. This number can be predicted by multiplying the areas of the chosen regions Ω_x and Ω_y, up to factors of 2π. Any error introduced by this finite approximation can be controlled, as finite matrices can be found arbitrarily close to our true kernel. The finite matrix's elements are found using (2.10). This requires a suitable set of basis functions $\{|b_i\rangle\}$, for which we have specific criteria.

Any kernel is related to a dual kernel. They have the same eigenvalues, and their eigenvectors are proportional inside the domains of interest (2.8). This will allow us to solve a reversed problem when doing so is simpler than the forwards one.

References

Aime, C., R. Soummer, and A. Ferrari (2002). "Total coronagraphic extinction of rectangular apertures using linear prolate apodizations". In: *Astronomy and Astrophysics* 389, pp. 336–344.

Roddier, F. and C. Roddier (1997). "Stellar Coronograph with Phase Mask". In: *Astro. Soc. of the Pacific* 109.1737.

Soummer, Rémi (2005). "Apodize Pupil Lyot Coronagraphs for Arbitrary Telescope Apertures". In: *ApJ* 618.1, pp. 161–164.

Slepian, D. and H. O. Pollak (1961). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — I". In: *The Bell System Technical Journal* 40.1, pp. 43–63.

Landau, H. J. and H. O. Pollak (1961). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — II". In: *The Bell System Technical Journal* 40.1, pp. 65–84.

Landau, H. J. and H. O. Pollak (1962). "Prolate spheroidal wave functions, fourier analysis, and uncertainty — III: The Dimension of the Space of Essentially Time- and Band-limited Signals". In: *The Bell System Technical Journal* 41.4, pp. 1295–1336.

Slepian, D. (1964). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — IV: Extensions to Many Dimensions and Generalized Prolate Spheroidal Functions". In: *The Bell System Technical Journal* 43.6, pp. 3009–3057.

Slepian, D. (1978). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — V: The Discrete Case". In: *The Bell System Technical Journal* 57.5, pp. 1371–1430.

Slepian, D. (1976). "On Bandwidth". In: *Proceedings of the IEEE* 64.3, pp. 292–300.

Simmons, F. J. and F. A. Dahlen (2006). "Spherical Slepian Functions and the Polar Gap in Geodesy". In: *Geophysical Journal International* 166.3, pp. 1039–1061.

Birman, M. S. and M. Z. Solomjak (1987). *Spectral Theory of Self-Adjoint Operators in Hilbert Space*. Mathematics and Its Applications (Soviet Series). Dordrecht, Holland: D. Reidel Publishing Company.

Schwartz, Laurent (1952). "Transformation de Laplace des distributions". In: *Comm. Sém. Math. Univ. Lund*, pp. 196–206.

Paley, R.E.A.C. and N. Winer (1934). *Fourier Transforms in the Complex Domain*. New York: American Mathematical Society.

Stein, E.M and G. Weiss (2016). *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton, NJ: Princeton University Press.

Papoulis, Athanasios (1968). *Systems and transforms with applications in optics*. McGraw-Hill Series in System Science. Malabar: Krieger.

Rowland, Todd. *L2 Space*. From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. URL: http://mathworld.wolfram.com/L2-Space.html.

Hogan, J. A. and J. D. Lakey (2012). *Duration and Bandwidth Limiting: Prolate Functions, Sampling, and Applications*. Applied and Numerical Harmonic Analysis. New York, New York: Springer Science+Business Media, LLC, p. 3.

Gosson, Maurice A. de (2011). "Hilbert–Schmidt and Trace Class Operators". In: *Symplectic Methods in Harmonic Analysis and in Mathematical Physics*. Basel: Springer Basel, pp. 185–203.

Weisstein, Eric W. *Matrix Norm*. From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. URL: http://mathworld.wolfram.com/MatrixNorm.html.

Tropp, Joel A. *An Elementary Proof of the Spectral Radius Formula for Matrices*. URL: users.cms.caltech.edu/~jtropp/notes/Tro01-Spectral-Radius.pdf.

Chapter 3

Application to the General APLC

Having established so many properties of Slepian problems in general, we now apply them to the construction of our arbitrary-geometry pupil apodization functions and the resulting propagation of light through the coronagraph. We will assume that the coronagraph is ideal, in that each plane is a perfect Fourier conjugate of the plane before it, without distortion. We generally discuss the coronagraph as having four planes; the pupil, the (masked) image plane, the Lyot plane, and the instrument plane.

As discussed in § 1, the goal is to reduce the light from an on-axis source sufficiently to allow the far weaker planetary light to stand out. This must be done across the wavelength band that the camera is designed to register, and across the angular range in which we expect nearby planets to appear from Earth's location. It should also be relatively unaffected by perturbations to the incoming wavefront and slight axial misalignment. The reduction of the starlight can occur though either a direct absorption of the light, as on a solid mask or Lyot stop, or through destructive interference.

Regardless of the means, the transmission of the light is concerned with

one region in each of the successive focal planes. Each transmits to the next a modification of the light from its region of concern. The implication is clear: *the eigenfunctions of the Slepian kernel for parts of the coronagraph are the natural modes for describing the transmission of light.* The effects will then be contained within the changes of the coefficients of those modes, barring a modification which breaks the paradigm.

Our focus is on the eigenfunctions created from the pupil-mask Slepian problem. These are the functions desired for apodization, but we will discuss in § 3.2 how these will give us explicit functional forms for the Lyot plane fields. (Discussion of how these apodizations sum to produce the blank pupil is delayed to § 4.3.1.) While we will speculate on the use of the mathematics for end-to-end or Lyot stop–instrument plane propagation, such will remain undeveloped.

If the coronagraphic mask is formally infinite, we can consider approximation with a large but finite mask. The Airy pattern produced by a purely circular pupil will contain $1 - 10^{-n}$ of the power within the radius $\lambda/D_P \approx 2 * 10^{n-1}$, so a mask of diameter $40\lambda/D_P$ will contain about 99% of the light for that simple pupil. (A mask of $30\lambda/D_P$ contains 98.5%; $20\lambda/D_P$, 98%; $10\lambda/D_P$, 96%.) Given the rapidity of the scaling, we will have to accept that any finite-mask approximations will necessarily contain a high margin of error.

To illustrate our points, we will sometimes use an explicit coronagraph of solid mask, width $\mathcal{N} = 5.0$. The pupil is circular, with circular central obstruction $R_S = 0.2$. The Lyot plane is set equal to the pupil.

§ 3.1 will show the general layout of the instrument, discussing the different optical planes and their functions. It will introduce the notation for explicit forms of functions in those different planes and the coordinate systems in use. The optical layout is summarized in figure 3.3. Table 3.1 summarizes the coordinate systems, while table 3.2 summarizes the abstract and applied forms of the basis and eigenfunctions.

§ 3.2 applies the lessons of section 2.5 to this layout. We show that the optically reversed problem is simple, and amenable to the techniques we have discussed so far. We then expand on our argument that all finite-mask coronagraphs are most naturally described in the Slepian modes.

Section 3.3 shows that with a circular focal-plane mask, the Zernike functions $R_t^{|m|}(\rho)e^{im\varphi}$ and their pupil-plane conjugates $\frac{J_{t+1}(r)}{r}e^{im\theta}$ fulfill the requirements set out in § 2.6. This basis will prove to have a tremendous number of additional benefits. With the basis chosen, § 3.4 summarizes the algorithm for finding the Slepian modes.

Section 3.5 summarizes our major conclusions.

3.1 Instrument Layout and Notation

Figure 3.1 depicts the layout of an APLC, which we will use to develop our notation for functions in the different planes. The coordinates of use are summarized in table 3.1 and displayed on their respective planes in figure 3.3. Notation for the eigenfunctions and eigenvalues are shown in table 3.2.

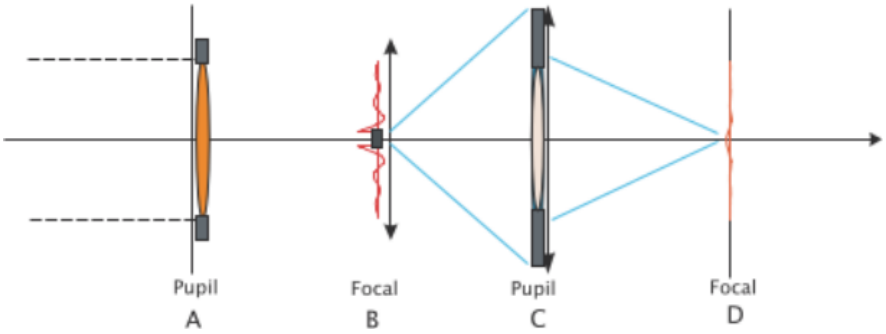


Figure 3.1: Layout of the general APLC. Figure taken from (Soummer et al., 2009).

3.1.1 Pupil plane

The coronagraph begins with the pupil plane, also referred to as plane A. It may or may not have a complicated shape. We refer to the distance between the optical axis and the outermost point which can transmit light as the pupil radius R_p. For circular pupils, this has the literal meaning. Distances perpendicular to the optical axis in the pupil plane, when measured in physical units such as meters, are denoted by $\mathbf{r}_1$. A non-dimensional distance measure will be introduced later; a summary for all planes is found in table 3.1.

The pupil is usually blocked in part by various structures, such as a central obstruction or supporting members for other parts of the instrument. We will

refer to the size of the central obstruction using R_S; if this is not a circular shape, then R_S is the largest radial distance in the shape. Collectively, these opaque areas in the pupil are referred to as the *secondary* structures. Figure 3.2 shows these for the irregularly-shaped JWST pupil.

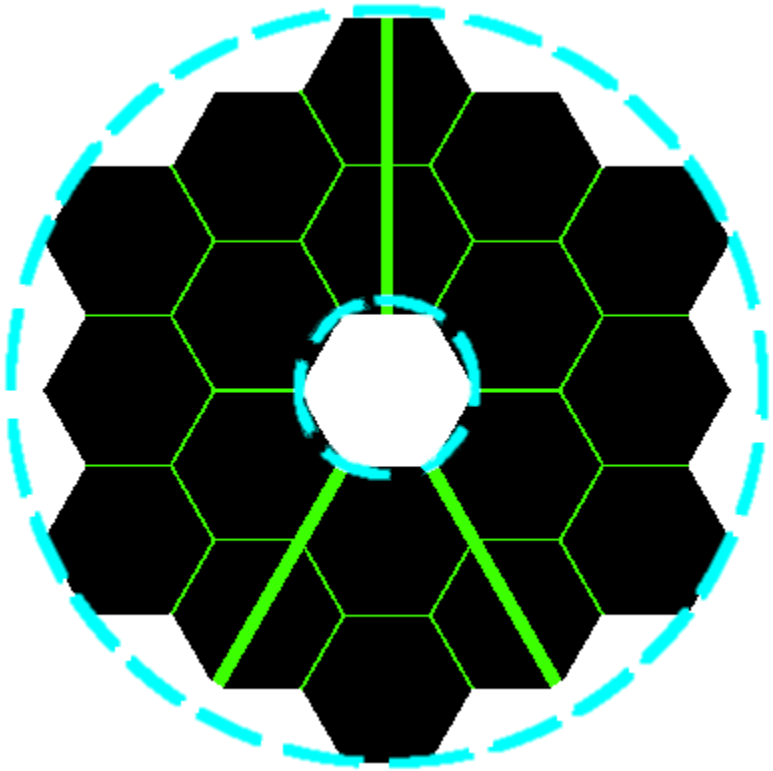


Figure 3.2: Illustration of R_P (cyan, outer circle), R_S (cyan, inner circle), and secondary structures (green) on the JWST pupil.

The lens or mirror of the pupil may have distortions which deform incident light rather than point-wise alter amplitude. We will assume that our pupil is distortion-free, and deal with any possible distortions in § 4.3. This lens or mirror transmits the light, with modification to amplitude caused by the apodization itself. It is possible, as noted in chapter 1, to introduce phase-altering apodizations. While we do not directly address such in our examples,

our formalism should handle this case.

3.1.2 Image plane

Assuming that our optical equipment is built to work in the paraxial regime, the focal plane of the pupil will be Fourier-conjugate to the pupil plane. We refer to this alternatively as the image plane, mask plane, or plane B. Unit-bearing distances from the optical axis are referred to by $\mathbf{r}_2$; again, a non-dimensional measure will be introduced later. We will neglect to write magnification and overall phase-factors introduced by the optics.

Directly on the optical axis itself is a mask, which intercepts some portion of the light. We will assume that the mask is isolated; that is, there are no support members which require notice. A solid mask absorbs that light falling on it, leaving only the surrounding radiation to carry on to the third plane. A phase mask introduces a π-phase shift, reversing the electric (and magnetic) field. Band-limited masks are themselves apodizations without any phase effects; other mask types, such as the four-quadrant phase mask, do change the phase of the incident light.

If we use M to symbolically be a function indicating the shape of the mask in this plane, then the radiation is therefore proportional to $(1 - M)$ for a solid mask and $(1 - M) - M = (1 - 2M)$ for the phase mask. More complex masks require either M to no longer be an indicator function, or to write these expressions as $(1 - M) + fM$ for $f(\boldsymbol{\rho})$ representing the alteration done by the mask. As any product of functions on the mask remains on the mask, the action of the mask on the electric field will have a linear operator

representation.

3.1.3 Lyot and Instrument planes

We again assume optical elements and arrangement so that the third plane is Fourier-conjugate to the second. This is the Lyot plane, also plane C. Distances are denoted with $\mathbf{r}_3$ when they have units. We will again ignore phase and magnification changes associated with the optics, including the inversion relative to the pupil plane.

The Lyot plane carries the Lyot stop, which causes further suppression of starlight through, as noted, removal of the Airy rings and central spot. The traditional Lyot stop consists of both an inner and outer Lyot stop, each perfectly circular. The inner stop is presumed to be a solid mask lying perfectly on the optical axis, blocking light inside its radius R_{L1}. The outer mask blocks all light lying beyond its radius R_{L2}. While R_{L2} is usually equal to or slightly smaller than R_P, to capture more of the diffracted starlight.

This circular shape is not required. It is common to consider a Lyot stop which is identical in shape to the pupil, as secondary structures from plane A often cause bright spots in plane C (Soummer et al., 2009). (We will demonstrate this fact in § 4.1.) A pupil-shaped Lyot stop will therefore block a significant additional amount of light compared to the circular one.

As products in one space become convolution in the Fourier-conjugate space, the Lyot plane's radiation is the convolution of the field which passed through the pupil, ϕ, and $1 - (1 - f)\widehat{M}$. ($\widehat{M}$ being the Fourier transform of the mask's shape.) If ϕ reacts so that $\phi \star (1 - f)\widehat{M} = \Lambda\phi$ – that is, the original

field is an eigenfunction under convolution with the mask's Fourier transform – then the total field in this plane is proportional to $(1 - \Lambda)$. The suppression of the on-axis starlight intensity therefore goes as $(1 - \Lambda)^2$. (For the simple $f = -1$ phase mask, this is usually written from the eigenvalues of $f \star \widehat{M}$ instead, resulting in $(1 - 2\Lambda)^2$ suppression.)

This suppressed starlight is transmitted to the fourth and final plane, the instrument or D plane. It is again Fourier-conjugate to the previous plane, and we usually neglect phase and magnification changes. Unit-carrying coordinates are denoted with $\mathbf{r}_4$.

We presume that the area of interest in that plane is an annular region lying between the inner working angle (IWA) and outer working angle (OWA). Our optical arrangement means that the distance $|\mathbf{r}_4|$ corresponds to an angular deviation from the optical axis. Since we are interested in angles which are $\lesssim 1$ arc-second, we can use the small-angle approximation when desired.

3.1.4 Non-dimensional coordinates

To better understand the optical behavior of the coronagraph, we should introduce a coordinate system to naturally align with the physics of the situation and the desired measurements. Consider the Fourier transform between

planes A and B,

$$\begin{aligned}
\widehat{f}(\boldsymbol{\rho}) &= \int \mathrm{d}^2 r_1 \, f(\mathbf{r}_1) \exp\left\{ i \frac{k}{L} \mathbf{r}_1 \cdot \mathbf{r}_2 \right\} \\
&= \int \mathrm{d}^2 r_1 \, f(\mathbf{r}_1) \exp\left\{ i \left(\frac{k R_P R_M}{L} \frac{\mathbf{r}_1}{R_P} \right) \cdot \left(\frac{\mathbf{r}_2}{R_M} \right) \right\} \\
&= \int \mathrm{d}^2 r_1 \, f(\mathbf{r}_1) \exp\left\{ i \left(\frac{\pi}{2} \left(\frac{D_M / L}{\lambda / D_P} \right) \frac{\mathbf{r}_1}{R_P} \right) \cdot \left(\frac{\mathbf{r}_2}{R_M} \right) \right\}
\end{aligned}$$

where $k = 2\pi/\lambda$ is the wavenumber of the incident radiation and L is the distance between the planes. We define

$$\mathcal{N} \equiv (D_M/L)/(\lambda/D_p) \tag{3.1}$$

as the angular width of the entire mask in units of λ/D_p. From this, we also define our normalization scale

$$\frac{R_P}{R_N} = \frac{\mathcal{N}\pi}{2} \tag{3.2}$$

The phase in the Fourier transform integrand is therefore $(\mathbf{r}_1/R_N) \cdot (\mathbf{r}_2/R_M)$, leading to natural definitions

$$r \equiv \frac{|\mathbf{r}_1|}{R_N} = \frac{\mathcal{N}\pi}{2} \frac{|\mathbf{r}_1|}{R_P} \tag{3.3}$$

$$\rho \equiv \frac{|\mathbf{r}_2|}{R_M} \tag{3.4}$$

We will use θ to indicate angles in the pupil (or Lyot) plane, and φ to indicate those in the image (or instrument) plane.

Since the Lyot plane is similar to the pupil plane, it is convenient to repeat the normalization and use $|\mathbf{r}_3|/R_N$. This will usually also be referred to with

r, except in cases of possible confusion.

The instrument plane is similar to the image plane, so it is possible to use $|\mathbf{r}_4/R_M|$ as before and refer to this distance as ρ. This is sometimes less useful as we are more interested in angular distances measured in units of λ/D_P. It is easy enough to convert using (3.1), and define

$$\zeta \equiv \frac{N}{2}\rho = \frac{|\mathbf{r}_4|/L}{\lambda/D_P} \tag{3.5}$$

as desired. Table 3.1 summarizes these nondimensional coordinate systems. Figure 3.3 displays them along a coronagraph.

Plane	**Distance**	**Definition**
Pupil	r	$r_1/R_N = \frac{\mathcal{N}\pi}{2} r_1/R_P$
Image/Mask	ρ	r_2/R_M
Lyot	r	$r_3/R_N = \frac{\mathcal{N}\pi}{2} r_3/R_P$
Instrument	ζ	$\frac{1}{2}\mathcal{N}\|\mathbf{r}_4\|/R_M = \frac{r_4/L}{\lambda/D_P}$

Table 3.1: Dimensionless distances in the different planes of the APLC. The instrument plane can also use a ρ-coordinate if it is convenient.

With these coordinates, we wish to establish expressions for the explicit forms of the basis and eigenfunctions for the pupil-mask mutual eigensystem.

The basis functions, which will be discussed more thoroughly in § 3.3, will be written in short form as $b_i(r,\theta)$ in the pupil plane. $\widehat{b}_i(\rho,\varphi)$ is their expression in the image plane, which will only be well defined on the mask. We will attempt to reserve i,j,k for situations with multiple basis functions written at once. After we establish that there exist radial and angular mode numbers, we will write them instead of i,j,k when the meaning is clear.

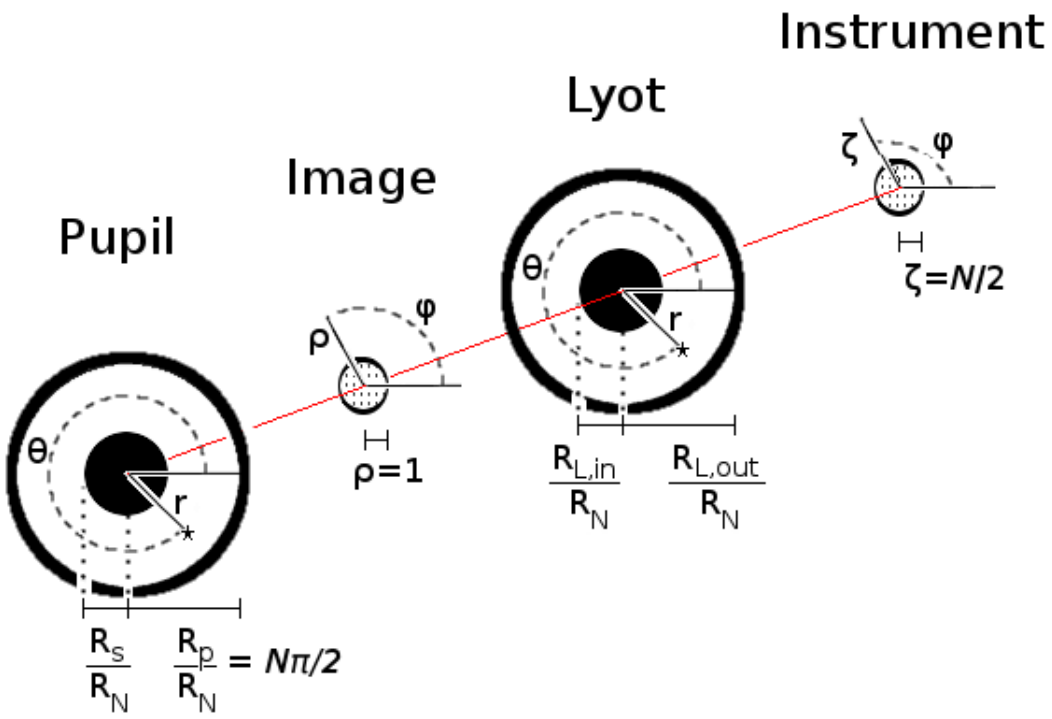


Figure 3.3: Layout of the coordinate systems on their respective planes, with reference to the common optical axis.

If we are referring to an eigenfunction in the pupil plane, we use the notation $\phi_a(r,\theta)$. The image plane will use the notation $\widehat{\phi}_a(\rho,\varphi)$. The label a means that this is the $a-$th eigenfunction, as ordered by descending eigenvalues Λ_a. The explicit components of the eigenfunctions in the chosen basis will be referred to by $V_{a,i}$, with conjugate $V^*_{a,i}$. We will attempt to reserve the letters a,b,c,d for such labeling when multiple vectors are being referred to at once. Equation (3.6) shows the explicit sum.

$$\phi_a(\mathbf{r}) = \sum_{tm} V_{a,tm}\, b_{tm}(\mathbf{r}) \tag{3.6}$$

With these coordinates, we may anticipate our $\sum_a \Lambda_a$, which we argued

in § 2.4 will be equal to the product of the area of the two regions $|\Omega_1|$ and $|\Omega_2|$ up to $(2\pi)^{-n}$. We have a two-dimensional problem; and, presuming a circular mask, its area is $|\Omega_2| = \pi(\rho = 1)^2 = \pi$. Once we calculate the area of our pupil in our scaled units, we immediately know that

$$\sum_a \Lambda_a = |\Omega_1|/4\pi \tag{3.7}$$

This will also be the number of eigenfunctions more concentrated inside the pupil than outside (i.e., $\Lambda_a > 1/2$ for $a \leq |\Omega_1|/4\pi$).

We remind ourselves that in such context we usually refer to $\sum_a \Lambda_a$ as the Shannon number. As such, it refers to the approximate size of the vector space (number of independent functions). This should therefore give us a lower bound on the number of basis functions necessary to describe our eigenfunctions.

Because of this, we use a circular approximation to form a first-pass approximation to the Shannon number. We do so as any pupil can be inscribed in a circle of radius $r = \mathcal{N}\pi/2$, as figure 3.4 demonstrates. This includes an estimation of a circular secondary inscribed within the actual central obscuration of the pupil, of radius R_S relative to the pupil radius. Taking this annular inscribing pupil as a (very rough) estimate of the area of the true pupil,

$$\sum_a \Lambda_a \lesssim \mathcal{N}^2 \left(\frac{\pi}{4}\right)^2 (1 - R_S^2) \approx \frac{5}{8}\mathcal{N}^2 \tag{3.8}$$

The scaling in $\mathcal{N}$ is the result of the two-dimensional nature of the area. Therefore, the number of eigenfunctions with $\Lambda_a \geq 1/2$ will scale as $\mathcal{N}^2$ for the full pupil as well.

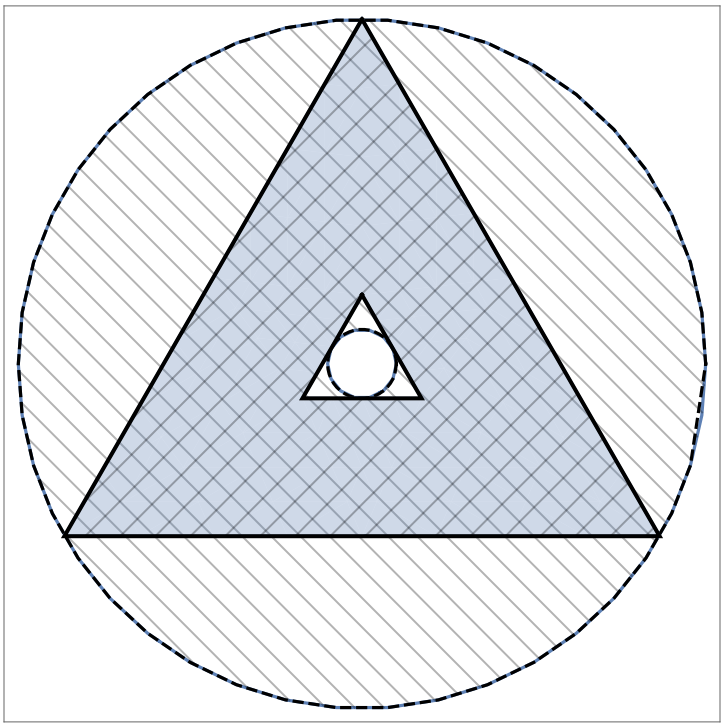


Figure 3.4: The inscription of an example pupil inside a circular one. This can be used for a crude estimation of the Shannon number, or for ensuring the creation of a positive apodization where desired (see 3.4).

3.1.5 Abstract expressions

Following the work of chapter 2, we now introduce notation in the abstract form. An eigenfunction or eigenvector will be denoted as $|a\rangle$, with complex conjugate for inner products denoted as $\langle a|$. Again, the label a means that this is the $a-$th eigenfunction, as ordered by descending eigenvalues Λ_a. Basis functions are labeled with $|i\rangle$ or $|tm\rangle$, where t will be a radial mode number and m an angular one. Their complex conjugate is $\langle i|$. The components of the eigenvectors are found by $\langle i|a\rangle$, which is the abstract alternative form for $V_{a,i}$.

We will use P_1 as the abstract version of the pupil indicator function, referring to areas where the pupil plane is open. P_2 will be the abstract mask

indicator function, referring to areas where the mask is located (i.e. $\rho < 1$). The kernel we will be dealing with is given by $K = P_2 P_1 P_2$; § 3.2 explains the reasoning behind this.

Our basis functions are naturally normalized on the mask: $\langle i| P_2 |j\rangle = \delta_{ij}$. The Slepian modes, therefore, are as well: $\langle a| P_2 |a\rangle = 1$, while $\langle a| P_1 |b\rangle = \Lambda_a \delta_{ab}$. This means that $\sum_i V^*_{a,i} V_{b,i} = \delta_{ab}$. We will need to remember to change the normalization of the functions as necessary to account for this when calculating certain physical quantities.

From 2.5, we remind ourselves that we can write the projection operators and kernel as

$$\begin{aligned} P_2 &= \sum_a |a\rangle \langle a| \\ P_1 &= \sum_a \frac{P_1 |a\rangle \langle a| P_1}{\Lambda_a} \\ K &= \sum_a \Lambda_a |a\rangle \langle a| \end{aligned} \tag{3.9}$$

The Lyot stop is more complex than the pupil and image planes. We will let P_3 denote the open areas of the Lyot stop. Since the pupil and Lyot planes are similar, both being Fourier conjugate to the image plane, we can denote the common open areas by multiplying the two projection operators $P_1 P_3$ or $P_3 P_1$. If the two are equal, then $P_1 P_3 = P_1^2 = P_1$. However, we must account for Lyot stops which do not match the pupil, $P_1 \neq P_3$.

In doing so, we have two different regions. The first is those areas of the pupil which are open, but are closed in the Lyot stop. An example would be

if the outer Lyot stop is undersized; these areas would be those between the outer edge of the stop and the pupil radius. In abstract terms, these are the regions $P_1 - P_1P_3 = P_1(1 - P_3)$. Since these are additional to the pupil plane, we refer to them as P_{1+}; note that $P_1P_{1+} = P_{1+}$.

The second region is the converse: areas open in the Lyot stop that are closed in the pupil plane. These would be secondary structures in the pupil that are not replicated in the Lyot plane. In abstract terms, they are $P_3 - P_1P_3 = P_3(1 - P_1)$. We will label these by P_{3+}; note that $P_3P_{3+} = P_{3+}$.

The relation between all four of these regions are therefore that $P_1 - P_{1+} = P_3 - P_{3+}$. A Venn diagram is shown in 3.5.

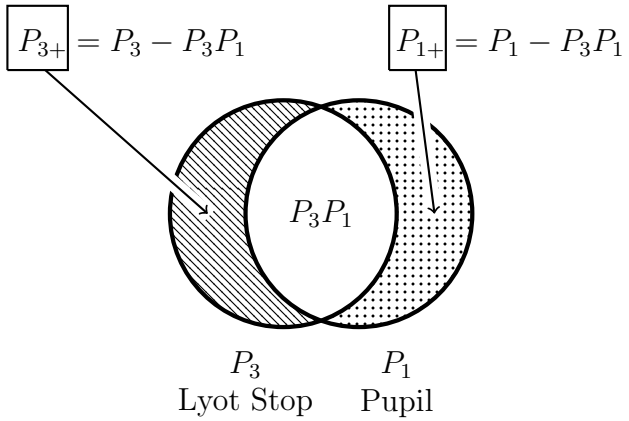


Figure 3.5: Venn diagram for the Lyot and pupil plane's different areas.

Table 3.2 summarizes the abstract and applied forms.

Symbol	Definition
$\|i\rangle$	$i-$th basis function
$\|tm\rangle$	basis function, radial mode t, angular mode m
$\|a\rangle$	$a-$th eigenfunction of the kernel
$\phi_a(r,\theta)$	Pupil and Lyot plane form of the $a-$th eigenfunction.
$\widehat{\phi}_a(\rho,\varphi)$	Image plane form of the $a-$th eigenfunction
Λ_a	$a-$th eigenvalue of the kernel
P_1	Abstract form, pupil (open area) indicator function
P_2	Abstract form, mask indicator function
P_3	Abstract form, Lyot stop (open area) indicator function
P_{1+}	Region of pupil not included in Lyot stop
P_{3+}	Region of Lyot stop not included in pupil

Table 3.2: Notation for the Slepian functions, basis functions, etc.

3.2 Optical Reversal for Finite Mask Applications

With the layout of the coronagraph done, we now turn to applying the work of chapter 2 to the problem of determining the Slepian modes of the pupil-mask kernel. Given the layout, our challenge is to find functions which are relatively concentrated on the mask, given that they must be defined solely in the pupil. This is especially challenging as we desire a simple set of starting basis functions to work in, "naturally" limited to the region of interest. Here, we wish to describe functions on the pupil, and so would hope for our natural basis to be defined there. Unless the geometry is restricted to the simplest case, this is impossible.

We therefore turn to the optically reversed problem. This would be the challenge of shining light through an opening the size of the mask, and attempting to determine the light from this which falls on the open areas of the pupil. We do so for all finite masks, ignoring their other optical transmission properties.

This optically reversed setup is exactly the Slepian dual, as discussed in 2.5, of the forwards problem of concern.

If we were to apodize this pseudo-window with any of the resulting Slepian modes, the Fourier transform would give us the pattern of the electric field in the open areas of the pupil. For the original problem, that "electric field" in the pupil would now be an apodization. Running from pupil to image plane, the pseudo-window's "apodization" is now the electric field which falls upon the mask. (There are some differences in normalization of

the field vs. the apodization, as discussed in 2.5, but they do not particularly matter at the moment.)

The amount of light from the pseudo-window backwards to the open areas of the pupil, relative to that through the pseudo-window, is the eigenvalue of the function. This is therefore also the relative amount of light which will fall on the mask, relative to that which would pass through the apodized pupil.

There are two major initial detractions to this dual problem. A non-constant mask will produce a significantly different field in the Lyot plane than was initially in the pupil plane, unlike the solid- and phase-mask cases. The formalism would seem, therefore, to only be of use in those two coronagraphic types. Moreover, we have argued in § 2.5 that the eigenfunctions of the kernel and its dual differ outside of the regions of interest. If the Lyot plane and pupil plane are different shapes, then would not using these eigenfunctions to calculate the fields be wrong?

The first objection may be met by the fact that both the original working functions and the eigenfunctions of the kernel will both be a complete basis, in the sense of § 2.2. That is to say, any function under consideration may be written as a weighted sum of either of these two sets. Complex masks may appear to produce functions which violate the restrictions we lay out in § 3.3, but on further consideration in § 3.3.2 and 4.1 we will show that this is not the case.

The second objection is handled as follows. Let our mask act on light entering the image plane, $\widehat{\Phi}$, by multiplying by a function f. The field in the Lyot plane is the sum of the transforms from $(1 - P_2)\widehat{\Phi}$, the region around the

mask, and $fP_2\widehat{\Phi}$.

This sum can be rewritten as $1\widehat{\Phi} + (f-1)P_2\widehat{\Phi}$. We now change our interpretation of it to read that the Lyot field comes the Fourier transform these two expression The first expression will clearly undo the Fourier transform, returning $\widehat{\Phi}$ to Φ. This was the light entering the image plane, and so is equal to the field from the pupil plane, entirely limited to the open areas of the pupil.

For the second piece, recall the definition of P_2: a projection onto the space of functions limited to the Fourier modes on the mask. These are, again, spanned by the eigenfunctions of our reversed kernel. It is these functions which are Fourier-transformed to produce part of the Lyot field. The fact that their values outside of the pupil's shape in the Lyot stop differ from the dual kernel's eigenfunctions in these locations is therefore irrelevant.

Both contributions to the field in the C plane are therefore, in principle, able to be written in terms of the Slepian modes of the reversed kernel of pupil-mask portion of the coronagraph. Any finite-mask coronagraph will therefore have its behavior described most naturally in these eigenfunctions.

We can also see this argument for completeness from the fact that we have shown the eigenfunctions of the *forwards* kernel look like $\Lambda_a^{-1/2}P_1\left|a\right\rangle$, (2.8). Since these form a complete basis for functions in the pupil, any light transmitted through the pupil can in principle be written using our basis. In practice, the stability of our numerical solutions and our desire to use a reasonably small number of modes will make the decomposition of some incident light impractical inside of the pupil. Frustratingly, this includes the off-axis light expected from a planet (see § 4.3).

While this case prevents us from writing the fields in the pupil and image planes, and from using this framework to its fullest, it does not prevent us from calculating fields in the Lyot stop. This is because, again, that field is the sum of the field in the pupil's open areas and the mask-limited effects. The field in the pupil's open areas is known even though in this scenario it is not as a decomposition over our modes. The second part is still the mask-limited portion of this, which is a small fraction of the total if our decomposition is so askew. The Lyot plane field is therefore still reasonable to calculate.

The fact that we have identified the natural modes of the pupil-image plane-Lyot plane portion of the system, and that we can always guarantee a reconstruction of the Lyot plane field, is joined by another great advantage of working with the reversed kernel. This problem relies on basis functions which are well-suited for the mask. Whereas the pupil is shaped by a great number of engineering concerns, the mask shape is in the hands of the coronagraphic teams. We are therefore free to select a *simple* shape whose basis functions obey as many of the nice properties of § 2.6 as possible.

While we will make some discussion on a rectangular mask in § 3.3.4, the properties we have found which result from a *circular* mask make it our complete focus.

3.3 Pupil-Mask Basis Functions and their Immediate Applications

We now must choose our working basis functions, based on the physical considerations of the coronagraph. To repeat, we will work primarily with a circular mask, as that is best suited to the coordinates described above and, as we will see, carries a large number of additional benefits. A brief aside on the use of rectangular coordinates and their basis functions, suited for a rectangular mask, is in § 3.3.4.

Since we want to describe realistic electric fields, the values must be finite everywhere on the mask. There are no other constraints on their values. We have chosen our normalizations such that $\rho = 1$ at the edge of the mask, and we desire an orthogonal family on that disk. Since this is a two-dimensional surface, our task will be simplified if we have a separation of coordinates. Having already chosen the circular mask, this will be the radial and angular modes.

We will choose the angular basis to be the simple $\frac{1}{\sqrt{2\pi}}e^{im\theta}$. This basis will require us to take care with complex conjugation, but is also naturally suited to describing the possible extensions to complex fields. We shall attempt to reserve the letters ℓ, m, n for angular mode numbers when we must write more than one at the same time and subscripts are unworkable.

For the radial basis we choose the Zernike polynomials Zernike, 1934

$$R_t^{|m|}(\rho) = \sum_{j=0}^{\Delta_{t,m}} (-1)^j \rho^{t-2j} \frac{(t-j)!}{j!(\sigma-j)!(\delta-j)!} \tag{3.10}$$

where $\Delta_{t,m} \equiv (t - |m|)/2$ must be a non-negative integer. The radial mode number obeys $t \geq 0$, and limits $-t \leq m \leq t$. We will attempt to reserve s and u as needed for other radial mode numbers where t' or t_1 would be hard to read. We will usually ignore writing the absolute value signs on R_t^m.

This (t, m) basis already fulfills many of the desirable properties that we laid out in §2.6. It is countable, with well-defined mode numbers. These polynomials have the great benefit that they are implicitly defined only for $\rho \leq 1$, and we will assume that $R_t^m(\rho > 1) = 0$ without inclusion of a Heaviside theta function to ensure it. *They are therefore naturally confined to the mask.*

The Fourier transforms of these basis functions are their appearance in the pupil plane. Again fulfilling one of our desired features from § 2.6, the Fourier transforms are analytic:

$$\mathfrak{F}[R_t^m(\rho)e^{im\phi}] \propto \frac{J_{t+1}(r)}{r}e^{im\theta}$$

(*NIST Digital Library of Mathematical Functions*, Eq. 10.22.64) where $J_{t+1}(r)$ is the Bessel function of the first kind and order $t + 1$:

$$J_\nu(z) = (z/2)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k!\Gamma(k + \nu + 1)}(z/2)^{2k} \tag{3.11}$$

For compactness' sake, we use the symbol

$$\mathcal{J}_{t+1}(r) = \frac{J_{t+1}(r)}{r} \tag{3.12}$$

and refer to these as the "jinc" functions by analogy to the $\frac{\sin x}{x}$ "sinc" functions. Conveniently, performing the Fourier transform on these functions naturally

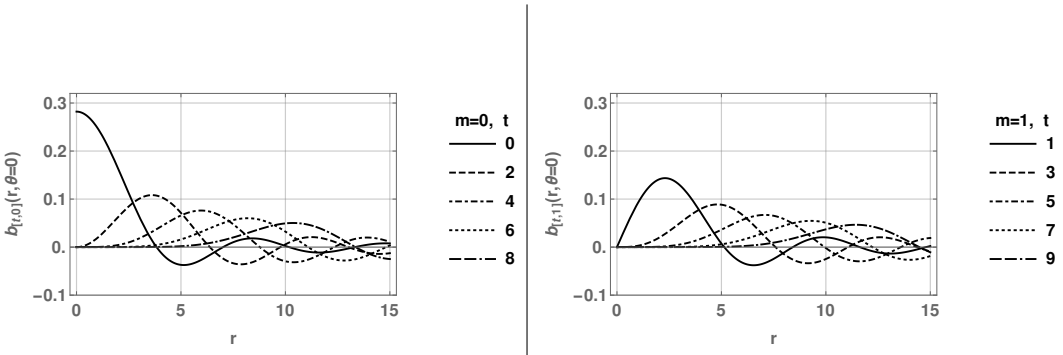
produces values of zero if $\rho > 1$, reinforcing our implicit cutoff at the mask edge.

We require that our basis functions be normalized as $\langle t, m | t, m \rangle = 1$. As the majority of our work will be done in the pupil plane, we move factors of (-1) from the Fourier transform to be in the image-plane expression of the basis function. This gives us our final form for these functions in both planes:

$$b_{t,m}(r,\theta) = \sqrt{2(t+1)}\frac{J_{t+1}(r)}{r} \cdot \frac{1}{\sqrt{2\pi}} e^{im\theta} \tag{3.13}$$

$$\widehat{b}_{t,m}(\rho,\varphi) = (-1)^{\Delta_{t,|m|}} \sqrt{2(t+1)} R_t^{|m|}(\rho) \cdot \frac{1}{\sqrt{2\pi}} e^{im\phi} \tag{3.14}$$

Examples are shown in figures 3.6 and 3.7.

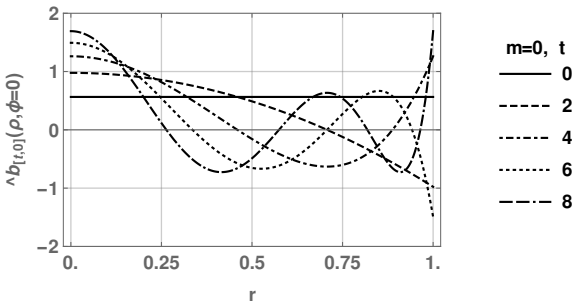


(a) Even radial modes **(b)** Odd radial modes

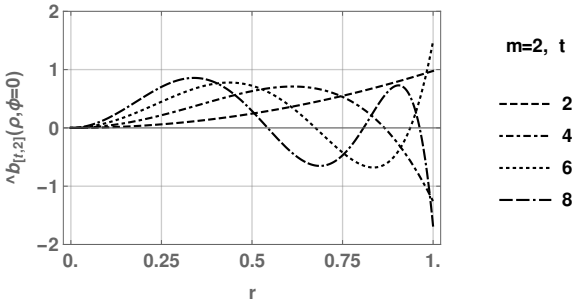
Figure 3.6: Examples of the radial parts of basis functions in the pupil plane.

Our kernel elements are then calculated via

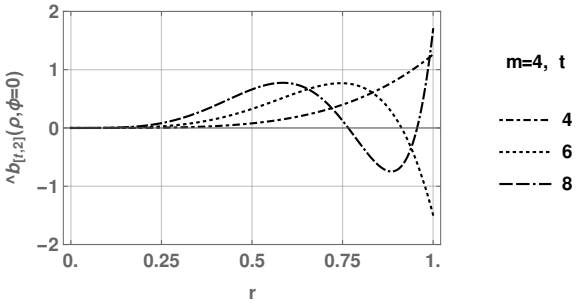
$$K_{t,m;t',m'} = \int \mathrm{d}^2 r \, P_1(\mathbf{r}) \, [b_{tm}(\mathbf{r})]^* \, b_{t'm'}(\mathbf{r}) \tag{3.15}$$



(a) Even t, $m = 0$ radial modes



(b) Even t, $m = 2$ radial modes



(c) Even t, $m = 4$ radial modes

Figure 3.7: Examples of the radial parts of basis functions in the image plane. All are shown for even t, but different values of m.

3.3.1 Choice and Implications of $t \leq T$

We now need to choose the maximum value for t and m to include in our work, and determine the size and type of the errors which will be introduced by this truncation. We aim for a relative error in the field of 10^{-5}, so that the error does not give us false contrast ratios larger than our usually desired 10^{-10}.

To begin, the Zernike polynomials are defined on $\rho \leq 1$ only for $t \geq |m|$, and require that $\Delta_{tm} \equiv \frac{t-|m|}{2}$ be a positive integer. We will discuss meanings and implications of this restriction more in 3.3.2; for now, we simply accept it.

If we have determined a maximum value for $t \leq T$, we automatically have a maximum possible value in $|m| \leq M \leq T$ as well. Presuming that we use all possible values of m, our bound on t means that we will have $\sum_{t=0}^{T} \sum_{m=-t;m++}^{t} 1 = \frac{(T+1)(T+2)}{2}$ basis functions to consider. If not, we have $\frac{1}{2}(2[T+1] - M)(M+1)$. The count for arbitrary cutoffs is shown in figure 3.8.

We now estimate lower bounds on the cutoff from coronagraphic parameters. From § 2.4, truncating the kernel as an $N \times N$ matrix will lead to errors of size Λ_N. The exponential decay § 2.4 of the eigenvalues past the Shannon number $\sum_a \Lambda_a = |\Omega_1|/4\pi$ is encouraging. Since this is the point where $\Lambda_a \approx 1/2$, a first method would be to use twice this number. Using the rough estimate at the end of 3.1.4 for the Shannon number, and again assuming all angular modes are used, $T \approx 1.5\mathcal{N}$. Going to three times the Shannon number gives $T \approx 2\mathcal{N}$. Only using up to M gives $T \approx \frac{5}{8(M+1)}\mathcal{N}^2 + M/2 - 1$ to $\mathcal{N}^2/(M+1) + M/2 - 1$.

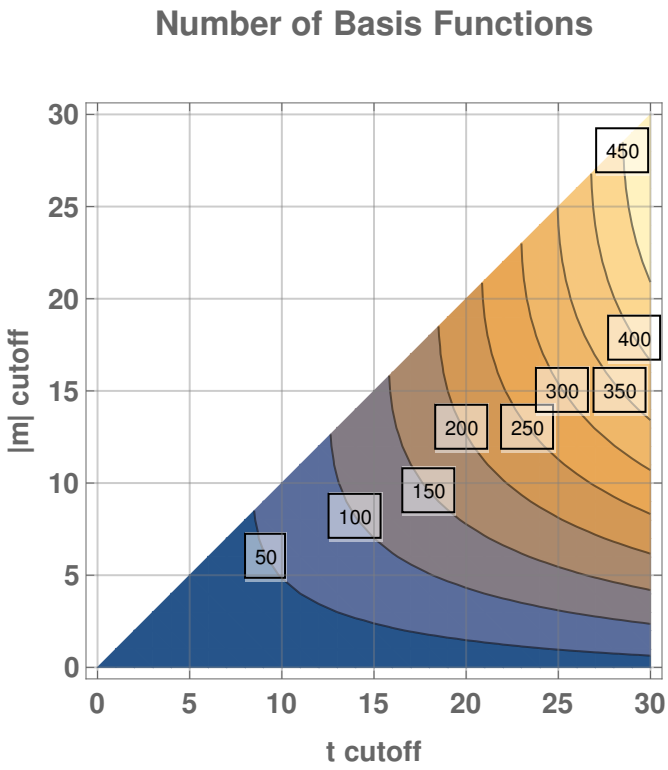


Figure 3.8: Number of basis functions given differing cutoffs in t and m.

A second method to estimate T relies on the fact that if the elements from the finite kernel excluding $T+1$ are small, then perturbation theory tells us that the deviations in the eigenvalues and eigenvectors will be of similar size to the excluded elements. Once all kernel elements involving a proposed T are calculated, we can determine their size relative to the largest value in the kernel so far. If they are below some desired tolerance, then it is reasonable to guess that the results will be of this tolerance.

Depending on the cost of calculating the kernel elements, it may be worthwhile to calculate the next values for the kernel to ensure that they are sufficiently smaller as to be acceptable. In the event that a faster method is desired, we can turn to the asymptotic behavior of the $\mathcal{J}_n(r)$ functions.

The basis functions $\mathcal{J}_n(r)$, as n becomes large, approaches the asymptotic form $\frac{1}{\sqrt{2\pi n}}\frac{1}{r}\left(\frac{er}{2n}\right)^n$, so that higher values of n are very rapidly suppressed. Because this asymptotic form is monotonic in r, the value of the function at the edge of the pupil is the maximum value which it will take over the entire pupil. An overestimate of the kernel entry is then to multiply the values the basis functions at the edge of the pupil by the area of the pupil.

Another prediction for kernel values is to use the ratio $\frac{b_{t+2,m}(\mathcal{N}\pi/2)}{b_{t,m}(\mathcal{N}\pi/2)}$ as an estimate for the kernel element ratio $\frac{K_{(s,n),(t+2,m)}}{K_{(s,n),(t,m)}}$. This prediction is done at the same m value, hence the $t+2$ instead of $t+1$; even and odd parts must therefore be done separately. Figure 3.9 shows an example; the estimated ratio approaches the true ratio very well around $t = 12$ or 14. Since this example is done for $\mathcal{N} = 5.0$, this value of t is closer to $2.5\mathcal{N}$ than the $1.5\mathcal{N} - 2\mathcal{N}$ predicted as the largest necessary t value from the Shannon number approach.

The actual values of some of those kernel elements for our standard example coronagraph, relative to the maximum in the kernel, are shown in 3.10. The relative error from excluding above $1.5\mathcal{N} \approx 8$ is of order $10^{-3.5}$, whereas above $2.5\mathcal{N}$ is of order 10^{-5}.

We will later show, in § 4.2 and § 4.3, that other uses of the framework encourage a higher T value, if the same tolerance requirements are extended there.

Figure 3.11 shows the change in the first ten eigenvalues Λ_a relative to their value at $T = 42$ for a broad range of T. The eigenvalues themselves are plotted in figures 3.12 and 3.13.

Once the relative change in Λ_1 crosses 10^{-15}, the matchine precision, the

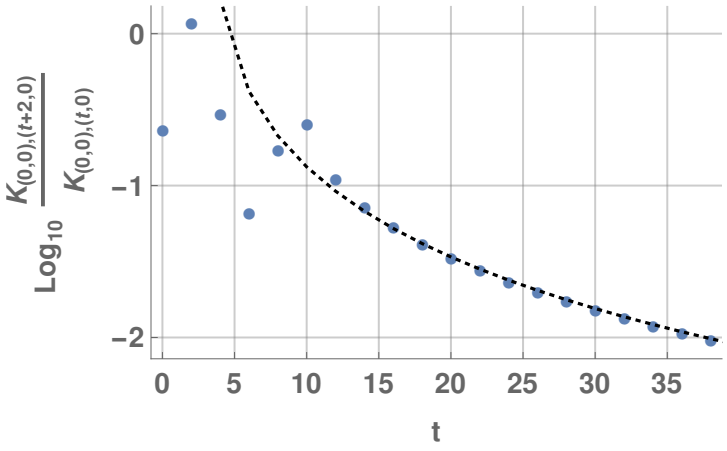


Figure 3.9: $\log_{10}|\cdot|$ of the ratio of subsequent kernel elements, at fixed $m = 0$. The dotted line is the ratio of successive basis functions at the edge of the pupil. The coronagraph's parameters were $\mathcal{N} = 5.0$, $R_S = 0.2$.

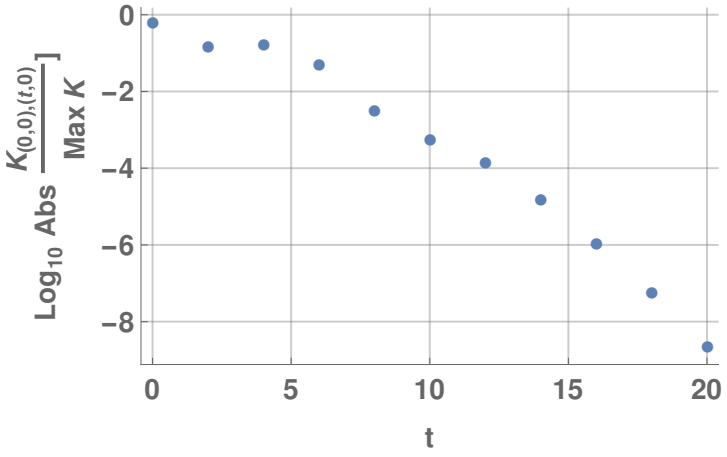


Figure 3.10: $\log_{10}|\cdot|$ of the first few kernel elements, relative to the largest, at fixed $m = 0$. The coronagraph's parameters were $\mathcal{N} = 5.0$, $R_S = 0.2$.

first six eigenvalues stabilize. This occurs at $T = 18$, about $3.5\mathcal{N}$. The seventh and eight cease changing by more than 10^{-6} by $T = 26$. The cases of $a = 9$ and $a = 10$, by contrast, are not settled down and exhibit wild fluctuations; however, they are themselves values below the machine precision, so this is not surprising.

This system was artificially limited to $m = 0$. If the full set of m had been used, then the eigenvalues which were of relatively high order $a = 5, 6$ would have been very far down ($a \sim 30$). While the general pupil will not have well-defined m numbers for the eigenfunctions, this still gives us a relative amount of confidence of the values of T needed to approach stability in eigenvalues of interest to us.

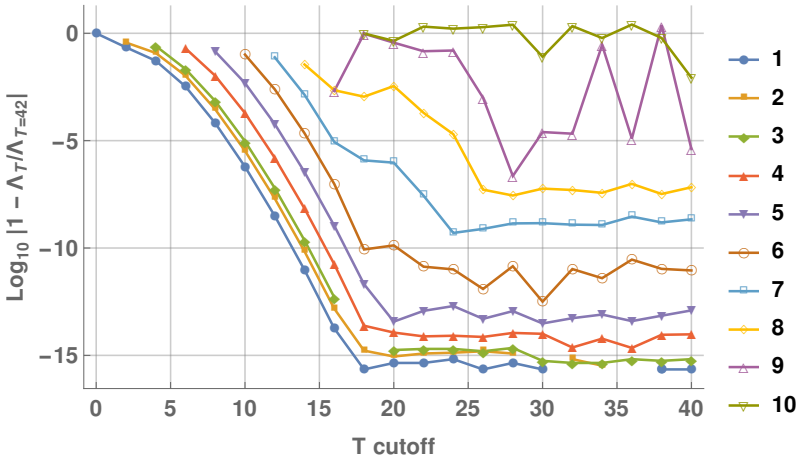


Figure 3.11: Relative difference between an eigenvalue for different cutoffs T and its calculated value at $T = 42$, at fixed $m = 0$. The coronagraph's parameters were $\mathcal{N} = 5.0$, $R_S = 0.2$.

3.3.2 Restrictions on (t,m)

The restrictions $m \leq t$ and $t - m = 2\Delta_{tm}$ for some integer Δ_{tm} require exploration, as we begin with neither the physical intuition for the effects nor practical knowledge of what occurs. We will use the Fourier integrals

$$\int_0^\infty \mathrm{d}r\ \mathrm{d}\theta\, r \frac{J_{t+1}(r)}{r} e^{im\theta} e^{ir\rho\cos(\theta-\varphi)} \times e^{i\ell\theta}$$

$$\int_0^1 \mathrm{d}\rho\ \mathrm{d}\varphi\, \rho R_t^{|m|}(\rho) e^{im\varphi} e^{ir\rho\cos(\theta-\varphi)} \times e^{i\ell\varphi}$$

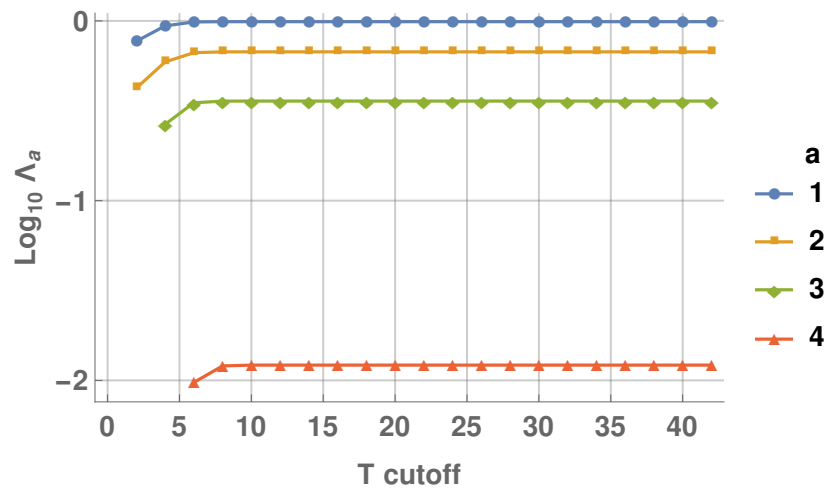


Figure 3.12: Different eigenvalues calculated using kernels of different cutoffs T, at fixed $m = 0$. The coronagraph's parameters were $\mathcal{N} = 5.0$, $R_S = 0.2$.

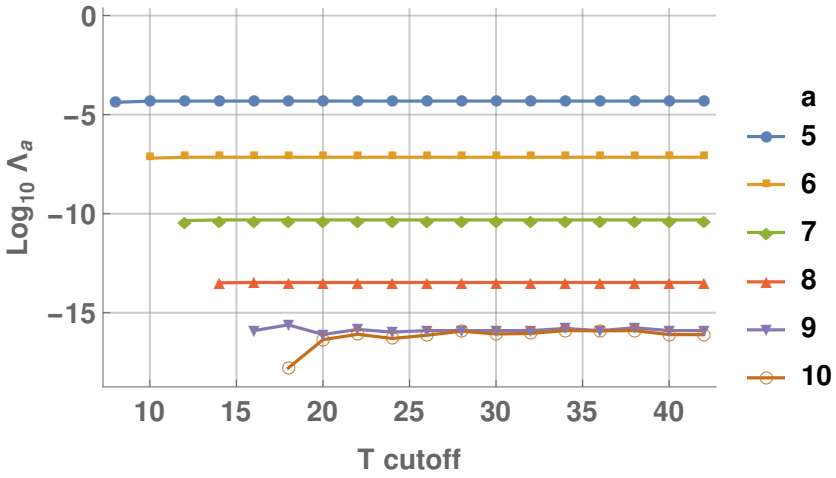


Figure 3.13: Different eigenvalues calculated using kernels of different cutoffs T, at fixed $m = 0$. The coronagraph's parameters were $\mathcal{N} = 5.0$, $R_S = 0.2$.

for different values of ℓ to naturally generalize the question. (In the event that $m > t$ in the image plane, we know that the Zernike polynomials are simply undefined there.) This also allows us to study the effect of mismatched $R_t^{|m|}(\rho)e^{i(m+\ell)\varphi}$. Such will occur when the focal-plane mask is anything other than a constant.

We break these into classes based on whether $|m+\ell| < |m|$, $|m+\ell| = |m|$, $|m| < |m+\ell| \leq t$, and $|m| \leq t < |m+\ell|$, and whether ℓ is even or odd. The integrals themselves we will call A or B, respectively, reflecting the plane in which they are carried out. The four classes are simply $1, 2, 3, 4$, with an e or o to show the parity of ℓ. In many cases, these integrals are zero or undefined, simplifying our exploration.

Integral A – Case 1e and 2: All we have done is change from the basis vector $|t, m\rangle$ to $|t, m+\ell\rangle$.

Integral A – Case 1o, 3o, and 4o: Performing the integral, we find the result is described by the sum of rational functions in ρ multiplying the elliptic $K(\rho^2)$ and $E(\rho^2)$ functions, for $\rho < 1$. If $\rho > 1$, it is the sum of rational functions of ρ times $K(\frac{1}{\rho^2})$ and $E(\frac{1}{\rho^2})$. This is true whether or not $t \geq m$ or $m \geq t+2$.

K is logarithmically divergent precisely at $\rho = 1$. Since we wish to consider finite electric fields, we exclude these functions from our consideration as being unphysical. This is despite the fact that they have perfectly finite integrals $\int_0^\infty \mathrm{d}\rho\, \rho |f(\rho)|^2$.

What if we were to place such an apodization on the pupil? In this case, the Fourier transform is no longer unlimited in r, and so the log divergence will not develop. It may be possible to attack this problem using the dilation operator 4.2 in ρ or $1/\rho$, as the integral $\int_0^R \mathrm{d}r\, J_n(r) J_k(r)$ is a known hypergeometric result. Regardless, the resulting functions are well-defined for all ρ, but we have not examined them any further.

Integral A – Case 3e and 4e: The integral in question is

$$\int \mathrm{d}^2 r \, \frac{J_{t+1}(r)}{r} e^{im\theta} \, e^{i\rho r \cos(\theta - \varphi)}$$

which, upon performing the angular integral leaves

$$\int_0^\infty \mathrm{d}r \, J_{t+1}(r) \, J_m(\rho r) \, e^{im\varphi}$$

We know that the radial integral will transform to the Zernike polynomial, $R_t^m(\rho)$, so long as $\rho < 1$ and our restrictions on (t, m) are obeyed. The integral will evaluate to zero for $\rho > 1$, an example of the Weber-Schafheitlin discontinuous integrals [(*NIST Digital Library of Mathematical Functions*) section 10.22]. A change of coordinates $r = r'/\rho$ leaves us with

$$\frac{1}{\rho} \int_0^\infty \mathrm{d}r' \, J_m(r) \, J_{t+1}\left(\frac{1}{\rho} r'\right) e^{im\varphi}$$

Therefore, if we obey the requirement $t - m = 2\Delta_{tm}$, but have instead that $m > t$, then the pupil-plane function transforms to

$$\frac{1}{\rho} R_{m-1}^{t+1}\left(\frac{1}{\rho}\right) e^{im\varphi}$$

for $\rho > 1$, and zero otherwise. These high-m modes can be used to describe the functions outside of the mask.

Physically, it means that the high angular mode $\frac{J_{t+1}(r)}{r} e^{im\theta}$ basis functions diffract light wholly outside of the $\rho < 1$ region when uninterrupted in r, a behavior we find non-intuitive. We have not explored the use of this phenomenon in a coronagraph as an alternative means of diverting starlight.

However, we can immediately see the similarity to the phase-vortex coronagraph (Foo, Palacios, and Swartzlander, 2005) and four-quadrant phase mask (Rouan et al., 2000), whose behavior we will touching on in § 4.1.

Integral B – Case 1e: In such a scenario, we are able to rewrite

$$R_t^{|m|}(\rho) = \sum_{j=0}^{[t-|m+\ell|]/2} c_j^{tm\ell} R_{t-2j}^{|m+\ell|}(\rho)$$

If we explicitly write out our polynomials, then we find that the c_j are the solutions to

$$\sum_{j=0}^{k}(-1)^{j+k}\frac{k!}{(k-j)!}\frac{(t-k-j)!}{(t-k)!}c_j^{tm\ell} =$$

$$\left(\frac{(\frac{1}{2}[t-|m+\ell|]-k)!}{(\frac{1}{2}[t-|m|]-k)!}\right)\left(\frac{(\frac{1}{2}[t+|m+\ell|]-k)!}{(\frac{1}{2}[t+|m|]-k)!}\right)$$

where $0 \leq k \leq \frac{1}{2}[t-|m|]$, and $\sum_j[\cdot] = 0$ for $\frac{1}{2}[t-|m|] < k \leq \frac{1}{2}[t-|m+\ell|]$. The Fourier transform is therefore the weighted sum of several well-defined basis functions of angular modes $m+\ell$ and radial modes $t, t-2, t-4, \ldots, |m+\ell|$. We have gathered a few examples in table 3.3 as a demonstration.

This result is part of the description of the behavior of phase-changing masks. We can see that as hoped, under these restrictions on ℓ they act as a linear operator in the space. They are therefore amenable to analysis in this framework.

Integral B – Case 1o, 3o, and 4o: The integrals evaluate to rational functions of r, multiplying either Bessel functions or the combination $J_1(r)\mathbf{H}_0(r) -$

(t,m,ℓ)	**Result**
$(7,7,-6)$	$\left(\frac{2}{5}\mathcal{J}_2(r) - \frac{2}{5}\mathcal{J}_4(r) + \frac{6}{35}\mathcal{J}_6(r) - \frac{1}{35}\mathcal{J}_8(r)\right) e^{i\theta}$
$(7,7,-4)$	$\left(\frac{2}{3}\mathcal{J}_4(r) - \frac{2}{7}\mathcal{J}_6(r) + \frac{1}{21}\mathcal{J}_8(r)\right) e^{3i\theta}$
$(7,7,-2)$	$\left(\frac{6}{7}\mathcal{J}_6(r) - \frac{1}{7}\mathcal{J}_8(r)\right) e^{5i\theta}$
$(7,5,-4)$	$\left(-\frac{1}{5}\mathcal{J}_2(r) - \frac{2}{5}\mathcal{J}_4(r) + \frac{3}{5}\mathcal{J}_6(r) - \frac{1}{5}\mathcal{J}_8(r)\right) e^{i\theta}$
$(7,5,-2)$	$\left(-\frac{2}{15}\mathcal{J}_4(r) - \frac{4}{5}\mathcal{J}_6(r) + \frac{1}{3}\mathcal{J}_8(r)\right) e^{3i\theta}$
$(7,3,-2)$	$\left(\frac{1}{15}\mathcal{J}_2(r) + \frac{4}{15}\mathcal{J}_4(r) + \frac{3}{5}\mathcal{J}_6(r) - \frac{3}{5}\mathcal{J}_8(r)\right) e^{i\theta}$

Table 3.3: Several results of the Fourier transform $\int \mathrm{d}^2\rho\, R_t^{|m|}(\rho)\, e^{i(m+\ell)\varphi}$, when $|m+\ell| < |m|$. Remember that $\mathcal{J}_n(r) = \mathcal{J}_{t+1}(r)$, so the angular and radial mode numbers are separated by an even number.

$J_0(r)\mathbf{H}_1(r)$, where $\mathbf{H}$ is the "Struve H-function" ((*NIST Digital Library of Mathematical Functions*) chapter 11). It is possible to Taylor expand these results in r, which results in a polynomial beginning at order $r^{|m+\ell|}$ and continuing without end. It is possible to use one of the intermediate formula from 3.3.3,

$$\left(\frac{z}{2}\right)^n = \sum_{j=0}^{\infty} \frac{((n+2j)(j+n-1)!)J_{n+2j}(z)}{j!}$$

to rewrite this sum as a legitimate series in basis functions. We neglect further study of odd ℓ.

Integral B – Case 2: As with integral A, we have at most only changed from $|t,m\rangle$ to $|t,\pm m\rangle$.

Integral B – Case 3e and 4e: Unlike the case $1e$, we cannot rewrite the Zernike polynomial as a finite sum. The result of the Fourier transform includes a sum of basis functions, but also factors of $[1 - J_0(r)]/r^2$. This would, at first glance, seem to mean that such an action breaks the system,

preventing phase-varying masks.

We therefore resort to Taylor expanding in ρ after integrating in φ. We have

$$\begin{aligned}
&\int \mathrm{d}\rho\, \rho \sum_{j=0}^{\Delta_{tm}} \sum_{k=0}^{\infty} \left(\frac{(-1)^j (t-j)! \rho^{t-2j}}{j!(\Delta-j)!(\sigma-j)!} \right) \cdot [r\rho/2]^{|m+\ell|} \left(\frac{(-1)^k (r\rho/2)^{2k}}{k!(k+|m+\ell|)!} \right) \\
&= \sum_{j=0}^{\Delta_{tm}} \sum_{k=0}^{\infty} \left(\frac{2}{r} \right) \left(\frac{r}{2} \right)^{|m+\ell|+2k+1} \left(\frac{(-1)^j (t-j)!}{j!(\Delta-j)!(\sigma-j)!} \right) \left(\frac{(-1)^k}{k!(k+|m+\ell|)!} \right) \\
&\quad \times \frac{1}{|m+\ell|+2k-2j+1} \\
&= \sum_{i,k=0}^{\infty} 2\mathcal{J}_{|m+\ell|+2[k+i]+1}(r) \cdot \left[\left(\frac{|m+\ell|+2[k+i]+1}{i!} \right) \right. \\
&\quad \left. \times \left(\frac{(-1)^k (|m+\ell|+2[k+i])!}{k!(k+|m+\ell|)!} \right) \sum_{j=0}^{\Delta_{tm}} \frac{(-1)^j (t-j)!}{j!(\Delta-j)!(\sigma-j)!} \right] \\
&= \sum_{n=0}^{\infty} \sqrt{2(|m+\ell|+2n+1)}\, \mathcal{J}_{|m+\ell|+2n+1}(r) \\
&\quad \times \left[\sqrt{2(|m+\ell|+2n+1)} \left(\frac{(|m+\ell|+2n)!}{n!(|m+\ell|)!} \right) \sum_{k=0}^{n} \frac{(-1)^k (|m+\ell|)! n!}{k!(n-k)!(|m+\ell|+k)!} \right]
\end{aligned} \tag{3.16}$$

The sum over k in the last expression is the ${}_1\mathrm{F}_1$ hypergeometric function, ${}_1\mathrm{F}_1(-n, a+1, 1)$, which we have left in explicit sum form.

As has happened before, this sum is only convergent when considered up to a maximum value in r. A few examples are in figure 3.14. Higher t converge faster, as do higher $|m+\ell|$. The figure indicating $(t, m, \ell) = (6, 4, 4)$, dotted line, shows the result of a sum whose coefficient was mistakenly calculated by

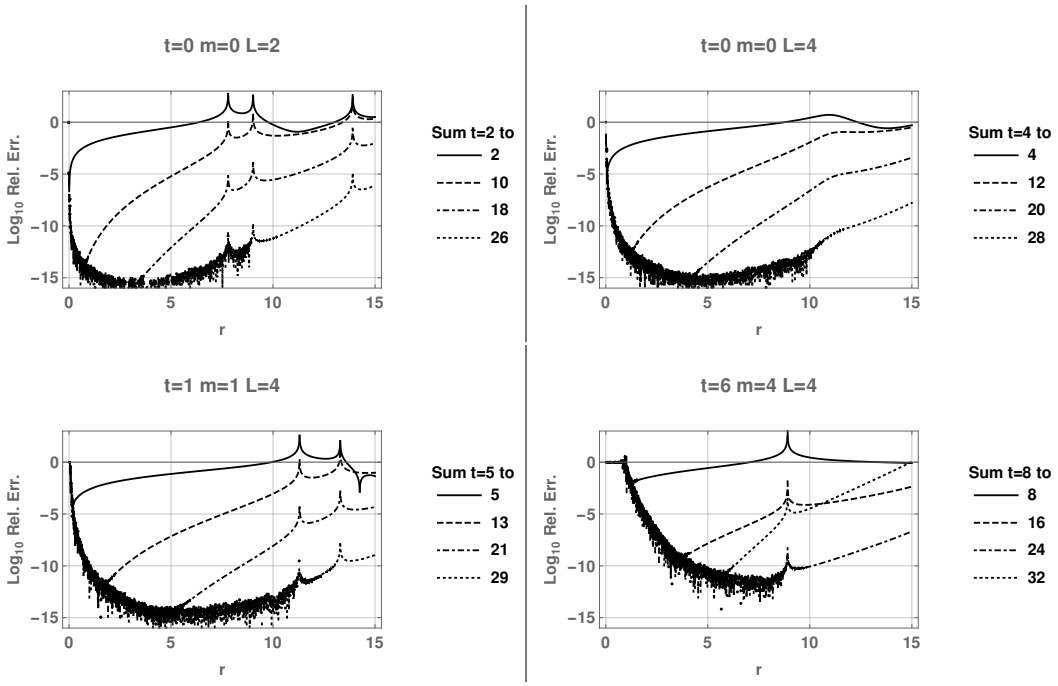


Figure 3.14: Convergence of the Fourier transforms of basis functions on the mask when multiplied by $e^{i\ell\varphi}$, with $|m| < |m + \ell|$.

a Taylor series expansion to insufficient order in r.

In summary: our choice of basis functions obey the restriction that $m \leq t$, as the Fourier transform from the pupil plane for $m > t$ are functions which are entirely off-mask, and so not of use in our Slepian kernel. The basis functions also obey $t - m = 2\Delta_{t,|m|}$ for an integer $\Delta_{t,|m|}$; if they did not, we would have electric fields with singularities in them, which we do not wish to consider. While a finite pupil mitigates this effect, we want a basis whose unrestricted Fourier transforms are finite in each plane.

We have also shown that on the mask, multiplication by a phase factor $e^{i\ell\varphi}$ does not break our choice of basis. Instead, when multiplying any of our chosen basis functions, the resulting Fourier transform to the Lyot plane is

expressible as a sum over the basis functions. The action of the image plane mask, even when not acting with constant phase, is still just a linear operator. It is best to act with an even ℓ. Variation in r and ρ are discussed in the second half of § 3.3.3.

3.3.3 Simple operators on basis functions

With our choice of basis functions, we can now write simple functions of r and ρ as linear operators. We will make note of those for which we have found use, though the study of those uses will be deferred to later sections.

Pupil-plane operators

We will start by examining the well-known recursion identities of the Bessel functions:

$$J_{\nu-1}(z) + J_{\nu+1}(z) = \frac{2\nu}{z} J_\nu(z)$$

$$J_{\nu-1}(z) - J_{\nu+1}(z) = 2\frac{\mathrm{d}}{\mathrm{d}z} J_\nu(z)$$

These give us immediate expressions for $r^{-1}\,|tm\rangle$ and $\partial_r\,|tm\rangle$.

$$\frac{\partial}{\partial r}\,|t,m\rangle = \frac{1}{2\sqrt{t+1}}\left(\sqrt{t}\,|t-1,m\rangle - \sqrt{t+2}\,|t+1,m\rangle\right)$$

$$\frac{1}{r}\,|t,m\rangle = \frac{1}{2(t+1)}\left(|t-1,m\rangle + |t+1,m\rangle\right)$$

The roots appear since we have set $\langle tm|tm\rangle = 1$. The angular derivative is simple,

$$\frac{\partial}{\partial\theta}\,|t,m\rangle = im\,|t,m\rangle$$

Despite the apparent successes, the $r-$involved expressions cannot be applied as it is. A single derivative clearly moves t by ± 1, leaving m unchanged. If we were applying these to $|t,|m| = t\rangle$ then, following the limitations of 3.3.2, we would be attempting to write a basis vector which is out of the space of functions we consider. There is also the challenge that the derivative is potentially ill-defined at $r = 0$, but as we expect this point to be excluded we are not concerned. A final problem occurs from the apparent $|-1,0\rangle$, which is not a function with finite squared integral.

If we consider even-powered combinations of these operators, they will return us to the correct (t,m) difference, but will have potentially dropped $t < m$. We will only examine the lowest even powers, $r^{-2}, r^{-1}\partial_r$, and ∂_r^2. Each by themselves leave the $t < |m|$ a possibility, and so we must discard them as individual operators which we can write as matrices. In combination as the Laplacian,

$$\begin{aligned}\nabla_r^2\,|t,m\rangle &= \frac{t^2-m^2}{4t(t+1)}\sqrt{\frac{t+1}{t-1}}\,|t-2,m\rangle - \frac{1}{2}\left(1+\frac{m^2}{t(t+2)}\right)|t,m\rangle \\ &\quad + \frac{(t+2)^2-m^2}{4(t+1)(t+2)}\sqrt{\frac{t+1}{t+3}}\,|t+2,m\rangle \end{aligned} \tag{3.17}$$

we find the pleasant surprise that if $t = |m|$, the $|t-2,m\rangle$ term is multiplied by zero and so does not spoil the matrix. This also applies to the

$t = 0$ term despite using the recursion identities to step through the fictitious $|t = -1, m = 0\rangle$ vector, so we may regard this matrix as wholly general. For $t = 1$, the only choice of $m = \pm 1$ means that the apparent $(t^2 - m^2)/\sqrt{t-1}$ singularity is actually zero, as hoped.

The pupil-plane Laplacian is therefore a valid operator which can be written in matrix form with our chosen basis. Any power of it (e.g. $[\nabla^2]^2$) will also serve as a valid operator. While we have not made use of this, we believe that it may find a place in adaptive optics.

We can use a repeated application of the recursion identities to determine the effect $r\,|tm\rangle$:

$$re^{\pm i\theta}\,|t, m\rangle = \sum_{j=0}^{\infty} 2\sqrt{t+1}\sqrt{t+2j+2}(-1)^j\,|t+2j+1, m\pm 1\rangle$$

For arbitrary r^n, we find it easier to derive our expression from the Taylor series for the Bessel function and equation 10.23.15 from (*NIST Digital Library of Mathematical Functions*).

$$J_{t+1}(z) = \left(\frac{z}{2}\right)^{t+1} \sum_{k=0}^{\infty} \frac{(-1)^k \left(\frac{z}{2}\right)^{2k}}{k!(k+t+1)!}$$

$$\left(\frac{z}{2}\right)^n = \sum_{j=0}^{\infty} \frac{((n+2j)(j+n-1)!)J_{n+2j}(z)}{j!}$$

$$\therefore z^n J_{t+1}(z) = 2^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(k+t+1)!} \left(\frac{z}{2}\right)^{2k+n+t+1}$$

$$= 2^n \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{(-1)^k (2(j+k)+n+t+1)(j+2k+n+t)!}{j!k!(k+t+1)!} J_{2(j+k)+n+t+1}(z)$$

$$= \sum_{p=0}^{\infty} J_{n+2p+t+1}(z) \left(\sum_{k=0}^{p} \frac{(-1)^k 2^n (n+2p+t+1)(k+n+p+t)!}{k!(p-k)!(k+t+1)!} \right)$$

A little algebra shows

$$\sum_{k=0}^{p} \frac{(-1)^k (k+n+p+t)!}{k!(p-k)!(k+t+1)!} = \frac{(-1)^p (n+p-1)!(n+p+t)!}{(n-1)!p!(p+t+1)!}$$

Therefore,

$$r^n e^{i\ell\theta} |t, m\rangle = \sum_{p=0}^{\infty} 2^n \sqrt{t+1} \left[\frac{(-1)^p \sqrt{t+n+2p+1}(n+p-1)!(t+n+p)!}{p!(n-1)!(t+p+1)!} \right]$$

$$\times |t+n+2p, m+\ell\rangle \tag{3.18}$$

assuming that $|\ell| \leq n$ and $(n - |\ell|)/2$ is an integer.

We note here that these expressions can be dangerous, as the coefficients grow without bound. They will converge for any finite r because the exponential decay of the $\frac{J_{t+1}(r)}{r}$ functions can eventually overpower the growth. This balance can make the expression dangerous to apply without caution.

The basis functions $|t, m\rangle$ necessary to apply this will extend further in t than the estimations supplied by 3.3.1. Convergence for the action of $r^n |0, 0\rangle$ and $r^n |10, 0\rangle$ is shown in 3.15. We can see that $|0, 0\rangle$ required more basis functions to converge acceptably. While the cutoff in t required to do so is very high, we note that this operator is blind to angular modes, and so if we

desire to study its effects we can increase T without increasing the cutoff in angular modes to match.

Use of these expressions are primarily for perturbations to wavefronts. Further study is therefore deferred to § 4.3.

Since we can now convert from polynomial to $\mathcal{J}_{t+1}(r)$ and back, we have the possibility of introducing an *algebra* on our basis functions. Referred to in § 2.2, an algebra a mathematical structure which converts a product of basis functions into a sum over basis functions:

$$|i\rangle \otimes |j\rangle = \sum_k c_k^{ij} |k\rangle$$

We draw on the additional identity

$$J_\mu(z)\, J_\nu(z) = \left(\frac{z}{2}\right)^{\mu+\nu} \sum_{i=0}^{\infty} \frac{(\mu+\nu+2i+1)!}{(\mu+\nu+i+1)!} \frac{(-1)^i}{i!(\mu+i)!(\nu+i)!} \left(\frac{z}{2}\right)^{2i}$$

from (*NIST Digital Library of Mathematical Functions*), 10.8.3. Following similar steps to the r^n derivation,

$$\frac{J_{t+1}(r)}{r} e^{im\theta} \times \frac{J_{s+1}(r)}{r} e^{in\theta} = \sum_{k=0}^{\infty} \frac{J_{t+s+2k+1}(r)}{r} e^{i(m+n)\theta}$$

$$\times \left[\frac{(t+s+2k+1)}{2} \sum_{i=0}^{k} \frac{(-1)^i}{(i+t+1)!(i+s+1)!} \frac{(t+s+2i+3)!}{(t+s+i+3)!} \frac{(t+s+i+k)!}{(k-i)!} \right]$$

(3.19)

Figure 3.16 shows the partial sums for a few example products. The series requires a very large number of terms to cause convergence. Even using up to

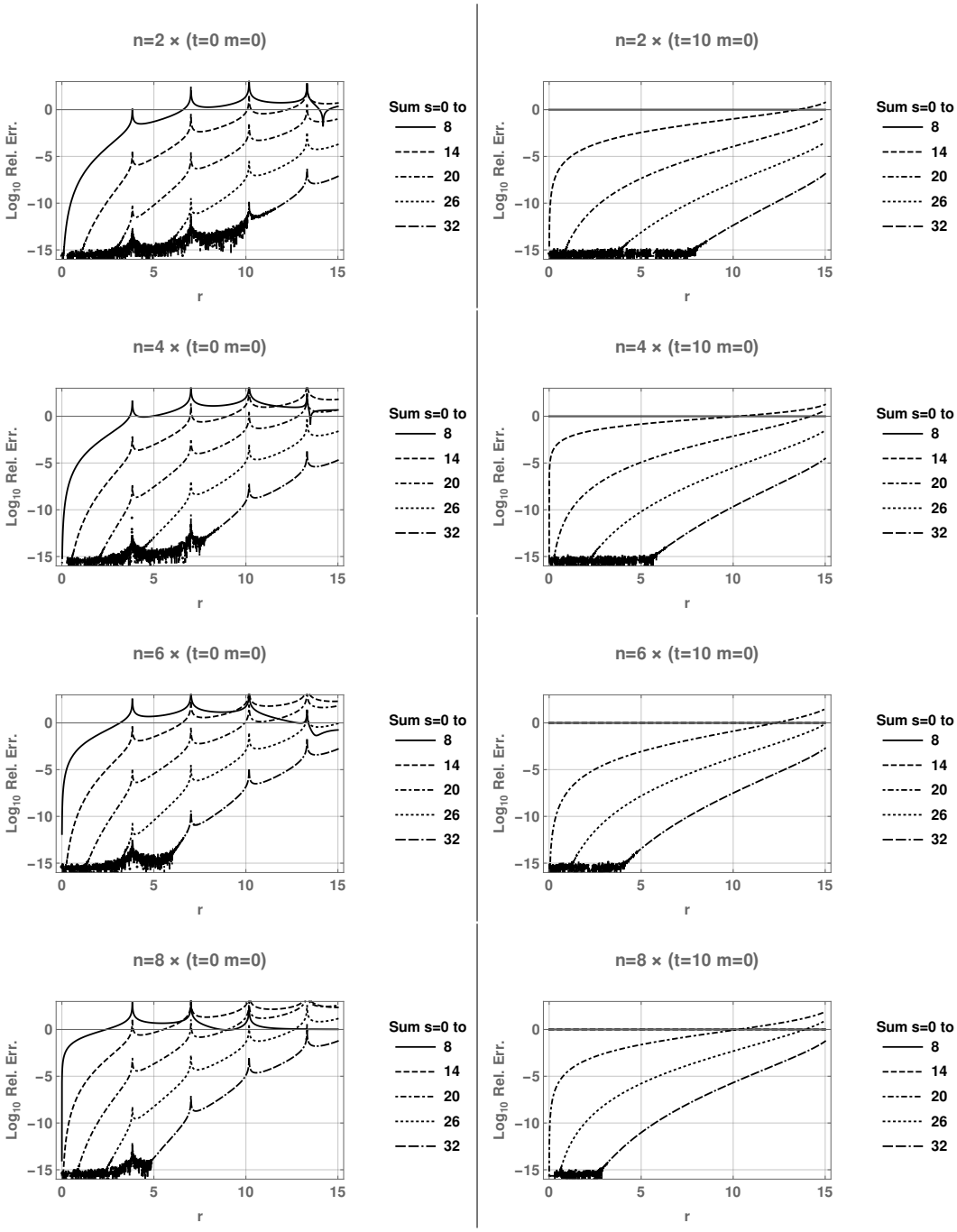


Figure 3.15: Convergence of the sums necessary to reproduce $r^n \left|t = 0, m = 0\right\rangle$ and $r^n \left|t = 10, m = 0\right\rangle$ for $n = 2, 4, 6, 8$.

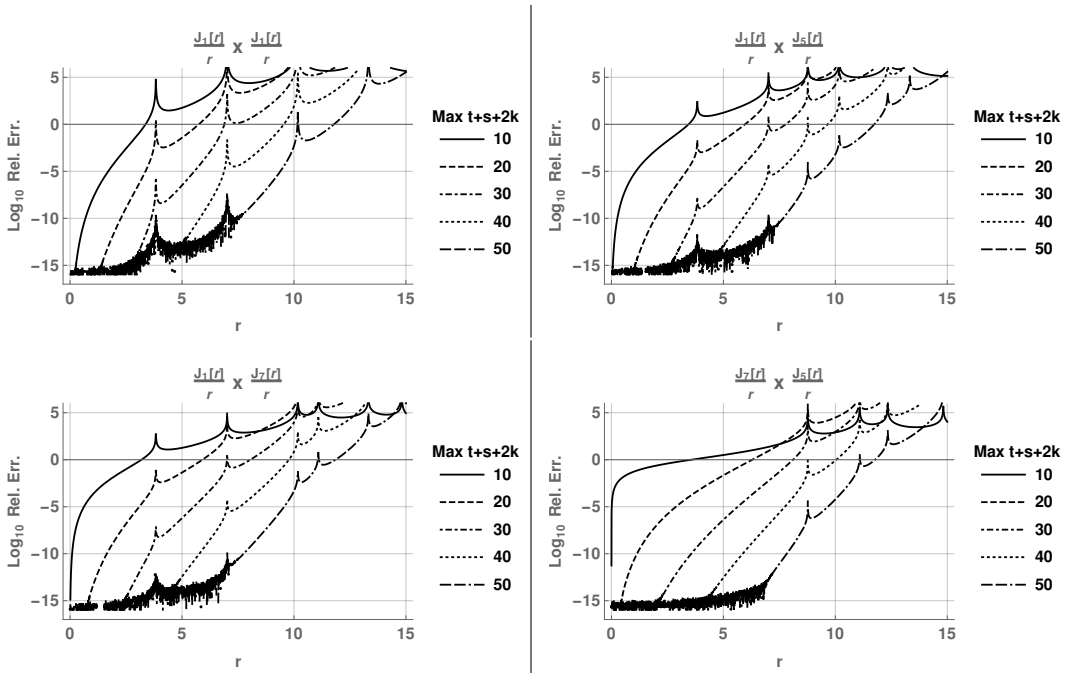


Figure 3.16: Convergence of the sums necessary to reproduce $\mathcal{J}_{t+1}(r) \times \mathcal{J}_{s+1}(r)$ for $t = 0, 4$ and $s = 0, 6$.

$t = 50$ this method cannot reproduce $[\mathcal{J}_1(r)]^2$ within the edge of a pupil for $\mathcal{N} = 6.4$. Given this, we consider the jinc algebra unlikely to be of general use.

One final operator formed of these combinations is $r\partial_r$. Following the recursive identities gives

$$r\partial_r |tm\rangle = t |tm\rangle + 2\sum_{j=1}^{\infty}(-1)^j \sqrt{(t+1)(t+1+2j)}\, |t+2j, m\rangle \tag{3.20}$$

Just as in the r^n case, it is an infinite sum. We can use this along with the ∂_θ operator to produce the components of a gradient $r\nabla$. We will not be able to find an eigenvector that serves as a zero for both the angular and radial modes; this merely gives us the resulting function.

By far the main use of the $r\partial_r$ operator is the creation of the dilation

operator, § 4.2. That will allow us to relate functions at different wavelengths with linear transforms and so create broadband results .

Image-plane operators

Since the image plane is conjugate to the pupil plane, $r \to \partial_\rho$ and $\partial_r \to \rho$ up to constant factors. This tells us that similar results for the image plane will hold; we should have even powers of ρ and ∂_ρ, or $\rho^n e^{i\ell\varphi}$, to create valid operators.

The Laplacian in ∂_ρ is straightforward. Its action on Zernike polynomials is known (Janssen, 2014),

$$\nabla^2_\rho \,|t,m\rangle = \sum_{\substack{s=|m| \\ \text{s++}}}^{t-2} \sqrt{(t+1)(s+1)}(t-s)(t+s+2)\,|s,m\rangle$$

Rather than discuss the individual actions of $\rho^n e^{i\ell\varphi}$, we will instead move directly to the product of basis functions. The result is also known (Tango, 1977); we adapt the explanation in (Haver and Janssen, 2013) to write:

$$\begin{aligned} R_t^{|m|}(\rho)e^{im\varphi} \times R_s^{|n|}(\rho)e^{in\varphi} &= \sum_u \left| C^{t/2,s/2,u/2}_{m/2,n/2,\ell/2} \right|^2 R_u^{|\ell|}(\rho)\, e^{i\ell\varphi} \qquad (3.21) \\ \ell &= m+n \\ C^{a,b,c}_{x,y,z} &= \sqrt{2c+1}(-1)^{a-b-z} \begin{pmatrix} a & b & c \\ x & y & -z \end{pmatrix} \end{aligned}$$

with the sum on u limited so that $\max(|m+n|, |t-s|) \le u \le t+s$. C is a Clebsch-Gordon coefficient; the matrix is a "Wigner 3j-symbol," and can be found in (*NIST Digital Library of Mathematical Functions*), chapter 34. This algebra is useful if the mask is apodized as in the bandpass coronagraph.

We have already discussed the behavior of multiplying by a pure phase on the mask in § 3.3.2. There we proved that while this action appears to move us into the realm of unacceptable $t < |m|$ modes, the Fourier transforms are still a sum over valid basis functions.

3.3.4 Basis functions for rectangular and other masks

We have so far only discussed the circular mask. Rectangular masks have been considered before (Aime, Soummer, and Ferrari, 2002), so we will at least consider some basics of these before returning strictly to circular masks.

Coordinates on the rectangular mask undergo separation of variables into ρ_x and ρ_y instead of (ρ, φ). We must separately scale these, which requires different $\mathcal{N}_x$ and $\mathcal{N}_y$. This means that x and y will require the appropriate handling in the pupil plane, in the same manner as r in 3.1.4.

Once done, we now desire basis functions which run from -1 to 1 and are orthonormal as $\int_{-1}^{1} \mathrm{d}\rho_x \, [f_i(\rho_x)]^* f_j(\rho_x) \propto \delta_{ij}$. The Legendre polynomials immediately suit this requirement, so our total basis functions in the image plane will be the product of Legendre polynomials $P_i(\rho_x)P_j(\rho_y)$.

The Fourier transforms of these are known (Fokas and Smitheman, 2012), and are generally of the form $(\sin x)/x^n$, $(\cos x)/x^n$, and combinations of these. Kernel elements are calculated with these transforms of the polynomials via equation (3.15). While we know that the Legendre polynomials possess useful recurrence relations, we do not know if these transforms do (and have not explored the matter). We therefore have not researched whether rectangular masks would have similar additional properties as the dilation

operator (*S* 4.2) and simple perturbation operators (§ 4.3).

Generalizing from the circular and rectangular mask, we can see that any mask whose shape is outlined by a separable coordinate system in two dimensions will follow this general procedure. In the event a mask shape is not from a separable system, we will likely have no simple basis functions and so our approach will fail. Numerical approximations for basis functions could be found by pixellating the mask plane, which may be less intensive than doing so for the pupil plane.

3.4 Algorithm for determining Slepian functions

We can now give an algorithm for calculating the Slepian eigenfunctions for an arbitrary pupil and circular focal-plane mask.

Begin by choosing the desired mask size $\mathcal{N}$, which is then used to scale a description of the pupil so that the edges are located at a radial distance $r = \mathcal{N}\pi/2$ from the center. Based on this scale, we can determine the number of eigenfunctions with $\Lambda_a > 1/2$ (aka the Shannon number) by dividing the area of the pupil $|\Omega_1|$ by 4π as given by (3.7). This area is bounded above by the circular case, $\pi(\mathcal{N}\pi/2)^2$, so the number of such eigenfunctions is no more than $\approx \frac{5}{8}\mathcal{N}^2$.

Assuming that all angular modes up to $|m| = t$ are included, cutting the radial modes off after a maximum value of $t = T$ results in $\frac{(T+1)(T+2)}{2}$ different basis functions. We absolutely must have more modes than the Shannon number. Ignoring later considerations for broadband behavior, perturbation study, and propagation past non-constant masks, we would recommend 2.5 to 3 times the Shannon number of basis modes be included. This will result in $T \approx 2\mathcal{N}$.

Once the basis functions are chosen, begin calculating the matrix elements via (3.15)

$$K_{tm;t'm'} = \int \mathrm{d}^2 r \, P_A(\mathbf{r}) \; [b_{tm}(r,\theta)]^* \; b_{t'm'}(r,\theta)$$

where the $b(r,\theta)$ are the pupil-plane versions in (3.13). We recommend only calculating unique pairings not related by complex conjugation. If estimation of the neglected kernel elements are above error tolerance (as described in

3.3.1), increase the number of modes and continue.

Form the resulting values into a proper matrix as needed, and then calculate the eigenvalues Λ_a and eigenvectors $V_{a,tm}$. The $a-$the apodization is then calculated by (3.6)

$$\phi_a(\mathbf{r}) = \sum_{tm} V_{a,tm}\, b_{tm}(\mathbf{r})$$

which has Λ_a of the energy focused onto the image plane mask relative to what passed through the pupil. This Λ_a is also the amount of energy which passes through the chosen pupil, relative to a very large, apodized, and unobstructed pupil.

The fact that the area of the pupil is equal to $4\pi \sum_a \Lambda_a$ provides a useful accuracy check. Very small eigenvalues are unlikely to converge completely to their true value, as are their eigenvectors. If they are needed, high precision must be used.

3.5 Summary

All coronagraphs of finite-sized focal plane mask are most naturally described by the Slepian eigenfunctions. This is true regardless of the real or complex nature of the apodization, and the real or complex action of the mask. The Slepian functions are the eigenvectors of the optically-reversed problem between the mask and the pupil. The kernel matrix for these eigenfunctions is $P_2P_1P_2$ (mask-pupil-mask). Fields in the Lyot plane can always be described with these eigenfunctions (in principle).

By scaling our pupil-plane coordinates as $r = (\mathcal{N}\pi/2)r_1/R_P$ and image-plane coordinates as $\rho = r_2/R_M$, we make it possible to use the Zernike polynomials $R_t^{|m|}(\rho)e^{im\varphi}$ as our working basis. Their Fourier transforms are the "jinc" functions $\frac{J_{t+1}(r)}{r}e^{im\theta}$. These have the desired properties laid out in 2.6. With these basis functions, we can create a finite matrix approximation for the kernel using (3.15).

The eigenvalues of the problem are the amount of power intercepted by the mask relative to the transmission by the pupil apodized by that mode. If placed in decreasing order their values decay exponentially towards zero. The sum of the eigenvalues is roughly equal to the number of modes with eigenvalue greater than one-half. This is known as the "Shannon number," and is considered to be the approximate size of the space of functions under consideration.

Lidskii's theorem (2.6) tells us that the Shannon number is equal to the area of the pupil (in r) divided by 4π. Rough circular approximation of the

pupil then means that this is $\sim 5\mathcal{N}^2/8$, and the $\mathcal{N}^2$ scaling will hold true for all pupil geometries. The Shannon number will also be equal to the trace of the matrix.

Future possibilities

We have left untouched a number of open avenues. One such is the possible construction of a kernel using the working area of the instrument plane, either between it and the Lyot stop or between it and the rest of the instrument back to the pupil plane. If $\mathcal{C}$ represents the operator propagating to the instrument plane and P_4 the restriction to the working region of that plane, then the forwards kernel would in theory be $\mathcal{C}P_4\mathcal{C}^\dagger$.

We speculate that the Slepian dual would be $P_4\mathcal{C}^\dagger\mathcal{C}P_4$, and that the annular Zernike polynomials (Dai and Mahajan, 2007) would serve as the natural basis. The modes of interest would be those which divert light away from the working region. This would require very high precision, as we would be interested in extremely low eigenvalue results.

It is likely possible to also create the Lyot stop–working region Slepian eigensystem. In the event that both are annular regions there may be explicit analytical formulae for the kernel elements as functions of the design parameters, as the annular Zernike polynomials would serve as a complete basis in both spaces. The analytical simplifications which occur (Slepian, 1964)s when both spaces are scaled copies of each other are ruled out if $IWA/OWA \neq R_{L,in}/R_{L,out}$. With such modes developed, the transfer from Lyot plane as developed by the pupil-mask Slepian modes in this thesis to the

instrument plane would in principle then be a matter of calculating overlap coefficients in the Lyot plane. In practice this may be too numerically intensive to be useful, a question we leave for future development.

References

Soummer, R., L. Pueyo, A. Ferrari, C. Aime, and A. Sivaramakrishnan (2009). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures, II. Theoretical Properties and Application to Extremely Large Telescopes". In: *The Astrophysical Journal* 695.1, pp. 695–706.

Zernike, F. (1934). "Diffraction Theory of the Knife-Edge Test and Its Improved Form, the Phase-Contrast Method". In: *MNRAS* 94, pp. 377–384.

NIST Digital Library of Mathematical Functions. "http://dlmf.nist.gov/, Release 1.0.14 of 2016-12-21". URL: "http://dlmf.nist.gov/".

Foo, Gregory, David M. Palacios, and Grover A. Swartzlander (2005). "Optical vortex coronagraph". In: *Optics Letters* 30.24, pp. 3308–3310.

Rouan, D., P. Riaud, A. Boccaletti, Y. Clénet, and A. Labeyrie (2000). "The Four-Quadrant Phase-Mask Coronagraph. I. Principle". In: *Publications of the Astronomical Society of the Pacific* 112.777, p. 1479. URL: http://stacks.iop.org/1538-3873/112/i=777/a=1479.

Janssen, A.J.E.M. (2014). "Zernike expansions of derivatives and Laplacians of the Zernike circle polynomials". In: *J. Opt. Soc. Am. A* 31.7, pp. 1604–1613.

Tango, William J. (1977). "The circle polynomials of Zernike and their application in optics". In: 13, pp. 327–332.

Haver, S. van and A. J. E. M. Janssen (2013). "Advanced analytic treatment and efficient computation of the diffraction integrals in the extended Nijboer-Zernike theory". In: *J. Europ. Opt. Soc. Rap. Public.* 8.13044.

Aime, C., R. Soummer, and A. Ferrari (2002). "Total Coronagraphic Extinction of Rectangular Apertures Using Linear Prolate Apodizations". In: *Astronomy and Astrophysics* 389.1, pp. 334–344.

Fokas, A. S. and S. A. Smitheman (2012). "The Fourier Transforms of the Chebyshev and Legendre Polynomials". In: URL: https://arxiv.org/pdf/1211.4943.pdf.

Dai, Guang ming and Virendra N. Mahajan (2007). "Zernike annular polynomials and atmospheric turbulence". In: *J. Opt. Soc. Am. A*. 1st ser. 24, pp. 139–165.

Slepian, D. (1964). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — IV: Extensions to Many Dimensions and Generalized Prolate Spheroidal Functions". In: *The Bell System Technical Journal* 43.6, pp. 3009–3057.

Chapter 4

Propagation with Slepian Modes

We have now demonstrated our ability to find the Slepian modes of the pupil-mask interaction. It naturally follows that we explore what simplifications and results appear in using these specialized modes to propagate through the coronagraph. All notation remain the same as in chapter 3. Figure 3.3 is repeated below, illustrating the generic layout we have.

We begin with § 4.1, discussing the propagation of a monochromatic plane wave. This will be done for an arbitrary mask, which we leave indefinite. We first do so with a single eigenfunction serving as apodization, then with a linear combination. We show here that the blank pupil, itself, can be written as a sum of Slepian modes, allowing this formalism to be applied to those coronagraphs without apodization. We write the resulting formulae for the fields and power at the different planes, and various figures of merit.

We then show in § 4.2 that our particular basis allows us to write a simple matrix operator, which represents a change of mask size or wavelength. We use this operator first in the abstract to study its application and general statements. Of great interest is the ability to calculate the derivative of the

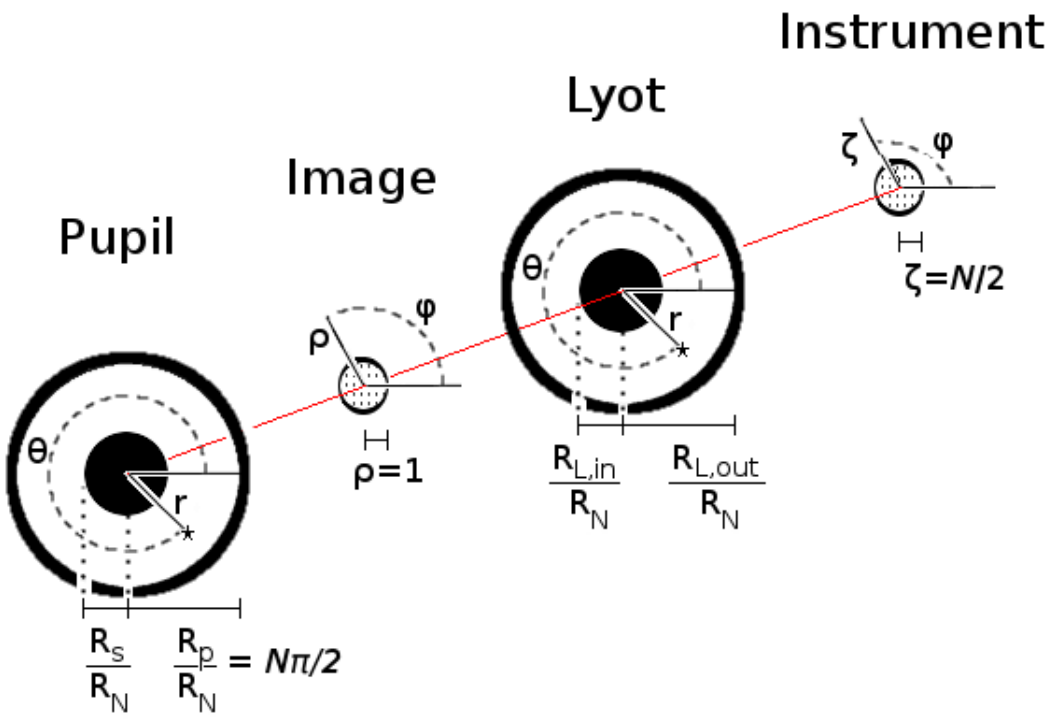


Figure 4.1: Layout of the coordinate systems on their respective planes, with reference to the common optical axis.

eigenvalues. This will not only allow estimation of achromaticity, but allows a general proof that all eigenvalues must increase with increasing $\mathcal{N}$. Next we show how to use the dilation operator to calculate results for incident starlight which does not match the design wavelength.

§ 4.3 discusses the handling of *non*-ideal wavefronts in the pupil plane. Our framework does well with perturbations which are slowly varying over the pupil plane. These include standard aberration functions, as well as finite-size effects for stars. The handling of far off-axis sources (planets) is then discussed. We conclude in § 4.4

4.1 Monochromatic propagation

Now that we have developed our method for generating the Slepian modes, we can step through the coronagraph to see what properties this reveals to us. Prior work has been purely numerical, so our semi-analytical approach represents a new method of analyzing the behavior of the systems.

We will act with an arbitrary $f(\rho, \varphi)$ on the mask. It will be easy to simplify in the cases where the pupil and Lyot stop are identical, and the APLC and phase mask cases. This gives us confidence in our general expressions. Note that this function f is wrapped in mask projection operators, $P_2 f P_2$, in order to determine the order in which it acts.

Various mask functions are shown in table 4.1. Our introduction to chapter 3 discussed how the vortex and bandpass masks can likely be approximated with large finite masks, and so fall under the scope of this thesis. Section 3.3.3 showed how to treat the product of Zernike polynomials or of pure phase functions as linear operators.

We only list one example of the band-limited masks from (Kuchner and Traub, 2002). These style of masks will likely be beyond our capability, as the authors describe them as being on the order of $\mathcal{N} \approx 1000 - 5000$.

This section will assume that the incident light is equal to the design wavelength for the coronagraph. Likewise, we must assume that the light falls normally onto the pupil (parallel to the optical axis). Both of these assumptions will be addressed later in the chapter.

We will start here by tracing the behavior through a pupil which has been

Mask name	f **Expansion**
Lyot	0
π phase	$-\delta_{ij}$
Sinc Band-limited	$1.217^{-1}[1-[\sin x]/x]$, $x \equiv \epsilon[\mathcal{N}\pi/2]\rho$
Vortex	$e^{i\ell\varphi}$
4QPM	$\sum_{\ell,odd} \frac{4}{i\ell} e^{i2\ell\varphi}$

Table 4.1: Representative examples of different types of focal plane masks.

apodized by a single Slepian function. We will then build from there to follow the results from a pupil that has been apodized by a sum $|\alpha\rangle = \sum_a \alpha_a |a\rangle$ of apodization functions. This includes the blank pupil, where $\alpha_a = \frac{\langle a|P_1|1\rangle}{\Lambda_a}$, $|1\rangle$ representing the constant function. (This expression draws on the logic from the end of 2.5 regarding proportionality of eigenvectors; it is derived in 4.3.)

At each plane we will give expressions for the field and its total power. We retain our $\int_{P1} \mathrm{d}^2 r\, \phi_a^2(\mathbf{r}) = \Lambda_a$ normalization for the field, rather than setting the maximum value of ϕ_a in the pupil to one (i.e. using ϕ as a transmission coefficient). The maximum value will instead be designated by $\mathcal{F}_a$. This means that all measurements of power are normalized by that coming through a hypothetical unobstructed, apodized pupil of sufficiently large size as to be called infinite.

We will also develop several figures of merit that are used. The *throughput* (aka the pupil photon throughput) refers to the number of photons which pass through the apodized pupil, as compared to an unapodized pupil (Soummer et al., 2009). If we instead compare the number of photons which reach the instrument plane, compared to those which pass through the apodized pupil,

we have the *residual* (Soummer et al., 2009). This is equal to a scaling of the energy inside the Lyot stop.

We also show that there are relatively simple expressions for the off-axis maximum of the PSF, which we approximate as the on-axis peak that results from removing the mask. For the combined apodization, we will also develop an expression for the instrument plane field behind the projection of the mask. The total power in a circular region of the instrument plane is also shown to have a relatively simple formal expression, though in practice it is not useful.

4.1.1 Single Slepian apodization

Let us choose the apodization $|a\rangle$ for our pupil; we assume that we therefore have the field inside the pupil $P_1 |a\rangle$. The total power in the plane, $|P_1 |a\rangle|^2$ is our familiar $\langle a| P_1 |a\rangle = \Lambda_a$, as $P_1^\dagger = P_1$ and $(P_1)^2 = P_1$.

In the **image plane**, the field just prior to the mask is $P_2 P_1 |a\rangle$, and just afterwards is $P_2 f P_2 P_1 |a\rangle$. The field around the mask in both cases is given by $(I - P_2)P_1 |a\rangle$. Since $|a\rangle = P_2 |a\rangle$ and the kernel is $K = P_2 P_1 P_2$, we can simplify the image plane field to

$$(P_1 - \Lambda_a) |a\rangle + \sum_b \Lambda_a f_{ab} |b\rangle \tag{4.1}$$

with the mask function expressed as

$$\begin{aligned} f_{ab} &= \langle b| f |a\rangle \\ &= \int_0^1 \mathrm{d}\rho \int \mathrm{d}\varphi\, \rho\, f(\rho,\varphi)\, \widehat{\phi}_a \left(\widehat{\phi}_b\right)^* \\ &= \sum_i \sum_j V_{a,i} \left(V_{b,j}\right)^* \int_0^1 \mathrm{d}\rho \int \mathrm{d}\varphi\, \rho f(\rho,\varphi) \widehat{b}_i \left[\widehat{b}_j\right]^* \end{aligned} \tag{4.2}$$

The power immediately after the mask in this plane is therefore

$$\Lambda_a(1-\Lambda_a) + \Lambda_a^2 \sum_b (f_{ab})^* (f_{ab})$$

originating from the region around the mask and through the mask, respectively.

We note here the difficulty with switching to a matrix notation $\underline{f}$, as this inner product does not go as $\underline{f}^\dagger \underline{f}$. This comes from the difference between matrix elements of the conjugate function, $(f^*)_{ab}$, and conjugates of the matrix element $(f_{ab})^*$, with the relation $(f^*)_{ab} = (f_{ba})^*$ or $f_{ba} = [(f^*)_{ab}]^*$. We will therefore retain explicit index notation, and write $f^*_{ab} = (f_{ab})^*$, with ${}^*f_{ab} = (f^*)_{ab}$ as needed.

For the **Lyot stop**, recall that the Fourier transform in these coordinates for this monochromatic case is the identity (2.3). This means that the abstract expression for the field we just derived, (4.1), is also the expression for the field just prior to the Lyot plane. Once the Lyot stop acts, the field inside is the restriction

$$P_3 \left(P_1 - \Lambda_a\right) |a\rangle + \Lambda_a \sum_b \left(f_{ab}\right) P_3 \, |b\rangle$$

The familiar results of the APLC ($f = 0$) and phase-mask ($f = -1$, $f_{ab} = -\delta_{ab}$) are simple to see. We can also see the action of f will reduce starlight when it acts to transfer light into the Slepian functions mostly concentrated outside the Lyot plane. These will primarily be the Slepian modes concentrated outside the pupil, as well, and therefore of small eigenvalue (high index a).

It will be useful to consider the field inside the Lyot stop using the division of the Lyot plane described in 3.1.5 and figure 3.5. We have the expression P_{1+} for regions of the pupil that lie outside the Lyot plane, such as areas blocked by undersizing the stop. The converse, P_{3+}, is for regions of the Lyot plane not inside the pupil, e.g. support structures.

With this notation, we have that $P_3 P_1 = P_3 - P_{3+} = P_1 - P_{1+}$. Any of these three expressions can be useful depending on which is easier to calculate with. This gives the Lyot stop field several useful expressions, each emphasizing some different aspect:

$$\begin{aligned}
&P_3 \left(P_1 - \Lambda_a\right) |a\rangle + \Lambda_a \sum_b \left(f_{ab}\right) P_3 |b\rangle \\
&\sum_b ([1 - \Lambda_a]\, \delta_{ab} + \Lambda_a f_{ab}) P_3 P_1 |b\rangle + \sum_b \Lambda_a \left(f_{ab} - \delta_{ab}\right) P_{3+} |b\rangle \\
&\sum_b \left([1 - \Lambda_a]\delta_{ab} + \Lambda_a f_{ab}\right) P_3 |b\rangle - P_{3+} |a\rangle
\end{aligned} \tag{4.3}$$

The first separates mask and off-mask contributions; the second explicitly displays the two disjoint regions in the Lyot plane; and the third groups all the effects before giving a correction due to pupil-Lyot stop mismatch.

The Lyot stop power, similarly, can have different useful forms.

$$
\begin{aligned}
&(1-\Lambda_a)^2 \langle a| P_3 P_1 |a\rangle + \Lambda_a^2 \langle a| P_{3+} |a\rangle \\
&\qquad + \Lambda_a^2 \sum_{bc} (f_{ab})^* f_{ac} \langle b| P_3 |c\rangle \\
&\Lambda_a (1-\Lambda_a)^2 + \Lambda_a^2 \sum_b \Lambda_b (f_{ab})^* f_{ab} \\
&\qquad + \Lambda_a^2 \left[\langle a| P_{3+} |a\rangle + \sum_{bc} (f_{ab})^* f_{ac} \langle b| P_{3+} |c\rangle \right] \\
&\qquad - (1-\Lambda_a)^2 \langle a| P_{1+} |a\rangle - \Lambda_a^2 \sum_{bc} (f_{ab})^* f_{ac} \langle b| P_{1+} |c\rangle \\
&(1-\Lambda_a)^2 \langle a| P_3 P_1 |a\rangle + \Lambda_a^2 \sum_{bc} (f_{ab})^* f_{ac} \langle b| P_3 P_1 |c\rangle \\
&\qquad + \Lambda_a^2 \langle a| P_{3+} |a\rangle + \Lambda_a^2 \sum_{bc} (f_{ab})^* f_{ac} \langle b| P_{3+} |c\rangle
\end{aligned}
\tag{4.4}
$$

The first separates different powers in f; the second, deviations from the case where the Lyot stop is equal to the pupil; the third, separate contributions from two regions of the Lyot plane. The appropriate expressions for APLC and phase mask follow on setting $f = (0, -1)$ and $P_3 = P_1$, as desired.

The **instrument plane** field is a Fourier transform of the Lyot stop field, and so its abstract expression does not change compared to (4.3). Likewise, the power which passes through the Lyot stop is the power arriving on the instrument plane, and so the power is equal to (4.4).

Since we have not constrained the Lyot stop, we cannot give a general functional expression for the instrument plane fields. In the event that $P_3 = P_1$ our situation is but little improved, as our basis functions were only defined

on the mask and are therefore unsuitable for calculating the working-area fields of interest. A general calculation for the instrument fields and intensity is therefore a numerical transform. We make a note on two points of possible interest at the end of this chapter.

We now turn to the expression for the throughput, the relative power (photon count, in monochrome) passing though the apodized pupil relative to the unapodized one. Again, we will call the maximum value of $|\phi_a|$ inside the pupil $\mathcal{F}_a$.

The power through the unapodized pupil is just the area of the pupil, $|\Omega_1|$. Recalling Lidskii's theorem (2.6), this area is related to the mask area and sum of eigenvalues by $|\Omega_1||\Omega_2|/(2\pi)^2 = \sum_a \Lambda_a$. In our scaled coordinates, the mask area is just π.

If we divide $|\phi_a|^2/(\mathcal{F}_a)^2$, we have our local measure of photon transmission. The total photon transmission is just the integral of this over the pupil, or $\Lambda_a/(\mathcal{F}_a)^2$. The throughput is then

$$throughput = \frac{1}{\mathcal{F}_a^2}\frac{\Lambda_a}{4\pi\sum\Lambda_a} \tag{4.5}$$

The determination of this $\mathcal{F}$ will be a computational challenge for efficient work in optimization. It is also a theoretical sticking point, as no simple expressions for it can be calculated (especially since it is the maximum inside the pupil, not the maximum overall).

The photon count at the instrument plane, relative to the number through the apodized pupil, cancels this $\mathcal{F}$ factor as it is common to both. The residual is therefore the same as the relative power ratio. We need only divide either

of the expressions from (4.4) by Λ_a. In the event that $P_3 = P_1$ and $f_{ab} = (0, -\Lambda_a\delta_{ab})$, we recover the known expressions $(1-\Lambda_a)^2$ and $(1-2\Lambda_a)^2$.

As our detection ultimately depends on the contrast between the starlight and planetary light, we seek a simplification for the expression of the peak intensity in the instrument field. We follow *REF* and approximate this with the peak of the on-axis field that results from removing the mask. For simple (APLC and phase) masks, we have found that this peak remains centered at $\zeta = 0$ so long as $\Lambda_a > 1/2$.

At this location, we enjoy the special circumstance that we are indeed behind the projection of the mask, and so may use the Zernike polynomials to describe the field. Since we are at zero, the only non-zero polynomials are those for which t is even and $m = 0$. These $R_t^0(0) = (-1)^{t/2}$, so our basis functions take on the value $\widehat{b}_{t,m=0}(0) = \sqrt{\frac{t+1}{\pi}}$.

The field itself is equal to $P_3P_1\left|a\right\rangle = (P_1 - P_{1+})\left|a\right\rangle$. The value of the field, $(\widehat{\phi}_a)_D(\rho = 0) = \left\langle\rho = 0\right|(P_1 - P_{1+})\left|a\right\rangle = \left\langle\rho = 0\right|P_2(P_1 - P_{1+})\left|a\right\rangle$ To find the value of the field, then,

$$\begin{aligned}(\widehat{\phi}_a)_D(0) &= \left\langle\rho = 0\right| P_3P_1\left|a\right\rangle \\ &= \left\langle\rho = 0\right| P_2(P_1 - P_{1+})\left|a\right\rangle \\ &= \left(\Lambda_a \sum_{t\,even} V_{a,t0}\sqrt{(t+1)/\pi}\right) - \left\langle\rho = 0\right| P_{1+}\left|a\right\rangle\end{aligned}$$

We can see a simple approximation with a correction factor due to regions of the pupil cut off by the Lyot stop. This correction factor is the average value of the eigenfunction over the region P_{1+}.

The approximate intensity peak is then

$$I_0 \approx \left| \left(\Lambda_a \sum_{t\,even} V_{a,t0} \sqrt{(t+1)/\pi} \right) - \langle \rho = 0 | \, P_{1+} \, | a \rangle \right|^2 \tag{4.6}$$

We may find expressions for the total power contained within a distance ζ. For clarity's sake, we will revert to using $\rho = \zeta/(\mathcal{N}/2)$ to describe distances in the D plane, and only treat the $f = 0$ APLC case. While our basis functions are only described for $\rho < 1$, we can still write general functions of ρ so long as we are careful. As we did not find this expression enlightening or useful, we only demonstrate it here for the APLC ($f = 0$).

The total power inside ρ_0 of the instrument plane is

$$\sum_{tm,t'm'} \int_0^{\rho_0} \mathrm{d}\rho\, \rho \int \mathrm{d}\varphi \int_{P3} \mathrm{d}^2 r \, (V_{a,tm})^* [b_{tm}(r,\theta)]^* (P_1(\mathbf{r}) - \Lambda_a)$$

$$\times \int_{P3} \mathrm{d}^2 r' \, V_{a,t'm'} b_{t'm'}(r',\theta') (P_1(\mathbf{r}') - \Lambda_a) \cdot e^{i\boldsymbol{\rho}\cdot(\mathbf{r}'-\mathbf{r})}$$

If we perform the ρ integral first, the exponential factor is the integrand and we have

$$\int_0^{\rho_0} \mathrm{d}^2\rho \, e^{\boldsymbol{\rho}\cdot(\mathbf{r}'-\mathbf{r})} = 2\pi i (\rho_0)^2 \mathcal{J}_1\left(\rho_0 |\mathbf{r}' - \mathbf{r}|\right)$$

We can separate the $|\mathbf{r}' - \mathbf{r}|$ in the jinc function by invokiong the Graf-Gegenbauer addition theorem (*NIST Digital Library of Mathematical Functions*) 10.23.7

$$\frac{J_\nu(|\mathbf{u} - \mathbf{v}|)}{z^\nu} = 2^\nu \Gamma(\nu) \sum_{k=0}^{\infty} (\nu + k) \frac{J_{\nu+k}(u)}{u^\nu} \frac{J_{\nu+k}(v)}{v^\nu} C_k^{(\nu)}\left(\frac{\mathbf{u}\cdot\mathbf{v}}{uv}\right) \tag{4.7}$$

where $C_k^{(\nu)}(\cos\alpha)$ is the Gegenbauer polynomial. Since in our case $\nu = 1$, the

Gegenbauer polynomial is just the Chebyshev U polynomial,

$$\begin{aligned} U_k(\cos\alpha) &= \frac{\sin\left([k+1]\alpha\right)}{\sin\alpha} \\ &= \sum_{\substack{j=-k \\ \text{j++}}}^{k} e^{ij\alpha} \end{aligned} \tag{4.8}$$

The power has therefore reduced to the sum of products of integrals

$$\begin{aligned} &4\pi \sum_{tm,t'm'} \sum_{k=0}^{\infty} \sum_{\substack{j=-k \\ \text{j++}}}^{k} (V_{a,tm})^* V_{a,t'm'}(k+1) \\ &\quad \times \sqrt{\frac{t+1}{\pi}} \left[\int_{P3} \mathrm{d}^2 r \, \frac{J_{k+1}(\rho_0 r)}{r} \frac{J_{t+1}(r)}{r} (P_1(\mathbf{r}) - \Lambda_a) e^{i(j-m)\theta} \right] \\ &\quad \times \sqrt{\frac{t'+1}{\pi}} \left[\int_{P3} \mathrm{d}^2 r' \, \frac{J_{k+1}(\rho_0 r')}{r'} \frac{J_{t'+1}(r')}{r'} (P_1(\mathbf{r}') - \Lambda_a) e^{-i(j-m')\theta'} \right] \end{aligned} \tag{4.9}$$

Whether this is more or less efficient that simply performing a numerical Fourier transform of the field from the Lyot plane will depend on the number of terms needed and the complexity of the integrals. We have defaulted to using numerical Fourier transforms and not explored this alternative. For larger ρ_0 we should be able to use the asymptotic form of the Bessel functions for easier numerical integration as needed. The exponential decay of the Bessel functions as a function of the order means that we can truncate the sum in a similar fashion to our determination of the cutoff in t, from section 3.3.1.

We have briefly examined expansions in ρ_0 using both the identity (*NIST*

Digital Library of Mathematical Functions) 10.23.1

$$J_\nu(\lambda z) = \lambda^\nu \sum_{k=0}^{\infty} \frac{(-1)^k(\lambda^2-1)^k(z/2)^k}{k!} J_{\nu+k}(z)$$

and Taylor expansion about $\rho_0 = 1$. The former simply created more complicated integrals to perform; the latter is related to the dilation operator, discussed in § 4.2, and rapidly becomes ineffective as ρ_0 increases past 1. Expanding about $\rho_0 = 0$ will give the same expressions as using the mask-projection fields.

The average power around a circle of ρ_0 is found by taking a derivative of this contained power with respect to ρ_0 and dividing by $2\pi\rho_0$. Doing so, the factor of $J_{k+1}(\rho_0 r)/r$ are converted to $2[J_k(\rho_0 r) - J_{k+2}(\rho_0 r)]$ when the derivative is acting on that integral. The averaged intensity is therefore

$$\begin{aligned}
&4 \sum_{tm,t'm'} \sum_{k=0}^{\infty} \sum_{\substack{j=-k \\ \text{j++}}}^{k} (V_{a,tm})^* V_{a,t'm'}(k+1) \\
&\qquad \times \sqrt{\frac{t+1}{\pi}} \left[\int_{P3} \mathrm{d}^2 r \; [J_k(\rho_0 r) - J_{k+2}(\rho_0 r)] \, \frac{J_{t+1}(r)}{r} (P_1(\mathbf{r}) - \Lambda_a) e^{i(j-m)\theta} \right] \\
&\qquad \times \sqrt{\frac{t'+1}{\pi}} \left[\int_{P3} \mathrm{d}^2 r' \, \frac{J_{k+1}(\rho_0 r')}{\rho_0 r'} \frac{J_{t'+1}(r')}{r'} (P_1(\mathbf{r}') - \Lambda_a) e^{-i(j-m')\theta'} \right] \\
&\quad + (\text{ other term in product rule}) \qquad (4.10)
\end{aligned}$$

Since neither the power interior to ρ_0 nor the angularly-averaged intensity at ρ_0 are simple expressions or seemingly useful calculations, we do not compute them for more complicated cases of the general pupil. § 5.1.1 will demonstrate the simplification which occurs to this expression when $P_{3+} = 0$

(the Lyot stop does not contain space which the pupil does not also contain).

4.1.2 Combined Slepian apodization

While satisfying, the behavior of a single Slepian mode does not give us free parameters to alter the eventual instrument-plane PSF, as is required for our planetary searches. We therefore turn to the behavior of sums of Slepian modes. Since we have demonstrated that the set of eigenfunctions forms a basis for functions in the open pupil, any function we desire can in principle be written in this way.

We will repeat the calculations just performed for a general apodization $|\alpha\rangle = \sum \alpha_a |a\rangle$. In functional form, we usually write this as $\Phi = \sum_a \alpha_a \phi_a$. For generic α, we prefer $\langle\alpha|\alpha\rangle = \sum |\alpha_a|^2 = 1$, as this is in line with our general normalization scheme $\langle a|a\rangle = 1$. Unless otherwise stated, we will assume that this is the case, and continue to use $\mathcal{F}_\alpha$ or plain $\mathcal{F}$ as the maximum value in the pupil.

We remind ourselves that the blank pupil, which we can abstractly represent as $|1\rangle$, can be written as an α sum with $\alpha_a = \frac{\langle a|P_1|1\rangle}{\Lambda_a}$. This reproduces a constant value of one in the pupil. (See § 4.3.2 for more detail). We will use a normalized $\langle\alpha|\alpha\rangle = 1$ version in this section; if a value of one is truly needed in the pupil, dividing $\frac{1}{\mathcal{F}}\alpha_a$ will provide it. This subsection therefore also allows us to follow the progression of unapodized coronagraphs.

The **pupil plane** field, as in the single-Slepian case, is just equivalent to $|\alpha\rangle$ if we continue to set the incident field intensity to one. The power in the pupil

is now

$$\langle\alpha|\, P_1\, |\alpha\rangle = \sum_a |\alpha_a|^2 \Lambda_a \tag{4.11}$$

relative to our unobstructed $|\alpha\rangle$ in the A plane.

In the **image plane**, the abstract field follows easily

$$\begin{aligned} &[(1-P_2) + P_2 f P_2]\, P_1\, |\alpha\rangle \\ &\sum_a \alpha_a \left[(P_1 - \Lambda_a)\, |a\rangle + \Lambda_a \sum_b f_{ab}\, |b\rangle \right] \end{aligned} \tag{4.12}$$

as does the power just after the mask

$$\sum_a |\alpha_a|^2 \Lambda_a (1-\Lambda_a) + \sum_{ac} \alpha_a^* \alpha_c \Lambda_a \Lambda_c \sum_b (f_{ab})^* f_{cb} \tag{4.13}$$

which divides neatly into to contributions around the mask and through it, just as in the previous case.

The abstract expression for the **Lyot plane** and **instrument plane** fields follow simply, and with the same multiple expressions as (4.3)

$$\begin{aligned} &\sum_a \alpha_a \left[P_3\, (P_1 - \Lambda_a)\, |a\rangle + \Lambda_a \sum_b (f_{ab})\, P_3\, |b\rangle \right] \\ &\sum_a \alpha_a \left[\sum_b ([1-\Lambda_a]\, \delta_{ab} + \Lambda_a f_{ab}) P_3 P_1\, |b\rangle + \sum_b \Lambda_a\, (f_{ab} - \delta_{ab})\, P_{3+}\, |b\rangle \right] \\ &\sum_a \alpha_a \left[\sum_b ([1-\Lambda_a]\delta_{ab} + \Lambda_a f_{ab})\, P_3\, |b\rangle - P_{3+}\, |a\rangle \right] \end{aligned} \tag{4.14}$$

The Lyot stop power (and, by extension, the instrument plane power) are now more complicated than the expressions (4.4), as we have the possibility of

interference between different modes.

$$\sum_{a,b} \langle a| P_3 P_1 |b\rangle \left[\sum_c \alpha_c^* \left(\delta_{ac}(1-\Lambda_a) + \Lambda_c [f_{ac}]^* \right) \cdot \sum_d \alpha_d \left(\delta_{db}(1-\Lambda_b) + \Lambda_d f_{bd} \right) \right]$$

$$+ \sum_{a,b} \langle a| P_{3+} |b\rangle \left[\sum_c \alpha_c^* \left(\Lambda_c [\delta_{ac} - (f_{ac})^*] \right) \cdot \sum_d \alpha_d^* \left(\Lambda_d [\delta_{bd} - f_{bd}] \right) \right] \tag{4.15}$$

An alternative expression can be found by remembering that $P_3 P_1 = P_1 - P_{1+}$ in the first line.

If $P_3 = P_1$ and we are in an APLC ($f = 0$), this expression reduces to

$$\sum_a |\alpha_a|^2 \Lambda_a (1-\Lambda_a)^2$$

whereas for the phase mask ($f = -1$) it reduces to

$$\sum_a |\alpha_a|^2 \Lambda_a (1-2\Lambda_a)^2$$

In either of these two simple cases, the power is a weighted sum of the individual modes.

We now turn to the throughput, residual, and off-axis peak for this arrangement. We will be using $\mathcal{F}$ to denote the maximum value of $|\Phi|$ which occurs inside the open areas of the pupil. This restriction is, again, non-analytic, and therefore will represent a major hurdle in the construction of optimization algorithms.

The throughput is the simple expression

$$\frac{1}{\mathcal{F}^2}\frac{\sum|\alpha_a|^2\Lambda_a}{4\pi\sum\Lambda_a} \tag{4.16}$$

If we rewrite this using (4.5),

$$\sum|\alpha_a|^2\left(\frac{\mathcal{F}_a}{\mathcal{F}}\right)^2(throughput)_a$$

we can see that this is a weighted sum of the individual mode throughputs. However, while $|\alpha_a|^2$ is less than one (by our general restriction $\langle\alpha|\alpha\rangle = 1$), we can not make the same guarantee for the $(\mathcal{F}_a/\mathcal{F})^2$ factor.

Our ability to improve the pupil's photon throughput depends on whether we are able to adjust the maximum of the weighted sum $|\alpha\rangle$ below the maximum of the principle components (those with large $|\alpha_a|^2$). This will result in a tendency for better throughput pupils to have flatter apodizations. If we cannot find such a combination, then the throughput will be strictly bounded by the throughputs of the individual modes.

The residual is the complicated expression (4.15) divided by $\sum_a|\alpha_a|^2\Lambda_a$. While we can extract pieces proportional to $|\alpha|^2(residual)_a$, the remainder cannot be qualified as positive or negative. We therefore neglect to rewrite the expression in general. For the simple APLC and phase mask cases, the expression will simplify to

$$\sum_a\frac{|\alpha_a|^2\Lambda_a}{\sum_b|\alpha_b|^2\Lambda_b}(1-\Lambda_a)^2$$

’ and

$$\sum_a\frac{|\alpha_a|^2\Lambda_a}{\sum_b|\alpha_b|^2\Lambda_b}(1-2\Lambda_a)^2$$

The individual mode residuals in these cases were $(1-\Lambda_a)^2$ and $(1-2\Lambda_a)^2$, which means that for the APLC and phase mask the residual in the Lyot plane is the weighted sum of the residuals of the individual modes. In these two cases, we cannot reduce the residual below that of the smallest one whose mode is involved in the sum.

We now construct the off-axis maximum for the PSF in the same fashion as for the single Slepian. We retain our assumptions that this maximum is the on-axis, no-mask maximum. We have found empirically that so long as $\sum_a |\alpha_a|^2 \Lambda_a \geq 1/2$, this occurs at $\zeta = 0$. This has not been rigorously tested, nor do we have a proof that it must be so.

So long as the peak truly is at the center, we may use the same behavior of the Zernike polynomials to simplify.

$$\begin{aligned}\widehat{\Phi}_D(\rho=0) &= \langle \rho=0|\, P_3 P_1\, |\alpha\rangle \\ &= \sum_a \alpha_a \left(\sum_{t\,even} \left[\Lambda_a V_{a,t0} \sqrt{(t+1)/\pi} \right] - \langle \rho=0|\, P_{1+}\, |a\rangle \right)\end{aligned}$$

As we may or may not be including Slepians whose peaks are not located at the center, we can not say that this is a linear combination of the peaks involved. The off-axis peak intensity in this approximation is just

$$I_0 = \left| \sum_a \alpha_a \left(\sum_{t\,even} \left[\Lambda_a V_{a,t0} \sqrt{(t+1)/\pi} \right] - \langle \rho=0|\, P_{1+}\, |a\rangle \right) \right|^2 \tag{4.17}$$

At this point we go further and develop an expression for the field value behind the projection of the mask. As we generically require a discrete Fourier transform for field values everywhere in the instrument plane, this will give

us a good control on the error introduced by the discretization. Moreover, if we desired we could take the inner product of this vector with itself to give us the power which is behind the mask projection in the instrument plane.

$$\begin{aligned}\Phi_{D,mask} &= \sum_a \alpha_a P_2 P_3 [(1 - P_2) + f P_2] P_1 \left|a\right\rangle \\ &= \sum_{ab} \alpha_a \Lambda_b ([1 - \Lambda_a]\delta_{ab} + \Lambda_a f_{ab}) \left|b\right\rangle \\ &\quad + \sum_{abc} \alpha_a \Lambda_a (f_{ac} - \delta_{ac}) \left\langle b\right| P_{3+} \left|c\right\rangle \left|b\right\rangle \\ &\quad - \sum_{abc} \alpha_a \Lambda_c ([1 - \Lambda_a]\delta_{ac} + \Lambda_a f_{ac}) \left\langle b\right| P_{1+} \left|c\right\rangle \left|b\right\rangle \end{aligned} \tag{4.18}$$

This requires us to calculate the inner products $\left\langle b\right| P_{1+} \left|c\right\rangle$ and $\left\langle b\right| P_{3+} \left|c\right\rangle$, which can be simplified by doing the calculations for the basis functions as $\left\langle j\right| P_{1+/3+} \left|k\right\rangle$. This is a subset of the matrix entry for the kernel, but we must perform the integral as there is no way to extract the value from the kernel. As a result, this expression may or may not be of use to us.

4.2 Broadband behavior

So far, all normalization has been carried out assuming that a single wavelength is of concern. We know this is not the case, and so seek relations between the behavior at different wavelengths. In doing so, we will show that there exists a special operator, which we call the *dilation operator* D_η. The existence of this operator in simple, closed matrix form is due solely to our choice of basis functions.

This matrix allows us to relate the kernels at different wavelengths with a simple transformation, and expand basis functions written at one wavelength in terms of the basis functions at another. In following this direction of thought, we are able to derive an expression for the derivative of the eigenvalues with respect to wavelength solely in terms of quantities at a single wavelength.

This matrix also allows us to follow the propagation of light through the coronagraph at different wavelengths, in the usual semi-analytical method.

4.2.1 The Dilation Operator

Recall our definition (3.3) $r = \frac{\mathcal{N}\pi}{2}\frac{r_1}{R_P}$ for $\mathcal{N} = (D_M/L)/(\lambda/D_P)$. A rescaling of the coordinate $r \to \eta r$ is equivalent to $D_M \to \eta D_M$ or $\lambda \to \lambda/\eta$; we will assume it is always the latter. This rescaling therefore describes broadband behavior.

The operator $r\partial_r$ is known to be the generator of rescaling, in the standard relation of Lie groups. In this context rescaling is generally referred to as dilation. This operator does not have any action on ρ, since $\rho = r_2/R_M$

will run from zero to one regardless of mask size. Our non-dimensional coordinates have placed all wavelength dependence in the pupil plane.

Below we show a standard proof of how the operator $r\partial_r$ is used to shift $r \to \eta r$ for any value of $\eta > 0$. In this derivation, we have relied on the earlier statement 2.1 that we are only considering functions with an infinite number of well-behaved derivatives. We have also implicitly assumed another characteristic of the function: that we may interchange the summations. [1]

$$
\begin{aligned}
f(\eta r) &= \sum_j \frac{\eta^j r^j}{j!} f^{(j)}(0) \\
&= \sum_j \left[e^{\ln \eta} \right]^j \frac{r^j}{j!} f^{(j)}(0) \\
&= \sum_j \left[\sum_k \frac{j^k (\ln \eta)^k}{k!} \right] \frac{r^j}{j!} f^{(j)}(0) \\
&= \sum_k \frac{(\ln \eta)^k}{k!} \sum_j \frac{j^k r^j}{j!} f^{(j)}(0) \\
&= \sum_k \frac{(\ln \eta)^k}{k!} \sum_j \frac{(r\partial_r)^k \, r^j}{j!} f^{(j)}(0) \\
&= \sum_k \frac{(\ln \eta)^k}{k!} (r\partial_r)^k f(r) \\
f(\eta r) &= e^{\ln \eta r \partial_r} f(r)
\end{aligned}
\tag{4.19}
$$

The exponential of the operator is to be meant as shorthand for the Taylor

[1] A series must be uniformly convergent for this to be true. The series definition of $J_\nu(z)$, (3.11), passes the Weirstrass M-test (*NIST Digital Library of Mathematical Functions*) §1.9(v) so long as the domain of z is limited, justifying our summation switch.

series, but it can be used as a well-behaved formal operator in algebraic manipulations. We will therefore call it the *dilation operator*

$$D_\eta = e^{\ln \eta \, r\partial_r} \tag{4.20}$$

$$\equiv \eta^{r\partial_r}$$

and introduce the occasional shorthand $\eta^{r\partial_r}$ for it.

There are several simple properties of D_η which follow immediately.

$$D_\eta D_{\eta'} = D_{\eta\cdot\eta'} \tag{4.21}$$

$$\left(D_\eta\right)^{-1} = D_{\eta^{-1}}$$

$$D_1 = \mathbb{I}$$

$$D_\eta r^n = \eta^n r^n$$

$$D_\eta \left[af(r) + bg(r)\right] = a\, D_\eta f(r) + b\, D_\eta g(r)$$

$$D_\eta \left[f(r)\, g(r)\right] = \left[D_\eta f(r)\right] \left[D_\eta g(r)\right]$$

The dilation operator itself in this basis can be found by taking the matrix exponential, shown in (4.20), of the $r\partial_r$ operator in (3.20).

$$\exp\left\{\ln(\eta)\, r\partial_r\right\}\; b_{t,m}(r,\theta) = b_{t,m}(\eta r, \theta)$$

In matrix form

$$D_\eta \left|tm\right\rangle = \sum_s [D_\eta]_{tm,sm} \left|sm\right\rangle \tag{4.22}$$

The matrix exponential is upper-triangular. Its entries $(D_\eta)_{t,s}$ are real polynomials in η of degree equal to s. We have not been able to find a closed-form expression for the explicit functions $(D_\eta)_{t,s}(\eta)$. However, Mathematica^{TM} is capable of performing the calculations for arbitrary η, so we are able to simply compute them once and store the results.

We may consider (4.22) purely as a mathematical identity in order to study the convergence. This is especially important as it may require us to include additional basis modes beyond those expected from 3.3.1.

While we show convergence behavior in 4.2.3, we will state an important conclusion here: we recommend that $\eta < 1$ be used to the greatest extent possible. This means that in analyzing a bandwidth, the smallest λ (highest $\mathcal{N}$) serve as the basis for the analysis. If necessary to go to shorter wavelengths, an $\eta = 1.1$ serves as a good upper limit.

4.2.2 Abstract application of the Dilation Operator

The equation (4.22) also serves as the relation equating a basis function written for one wavelength, to a sum of basis functions written for another:

$$b_{t,m}(\eta r) = \sum_s [D_\eta]_{t,s} b_{s,m}(r) \tag{4.23}$$

where we now interpret $\eta = \lambda_{old}/\lambda_{new}$. This will allow us to determine the effect of illuminating an apodization with light other than the design wavelength. In doing so, we will need to be careful with factors of η which appear in Fourier transforms from the differential $\mathrm{d}^2 r$.

Let's consider the kernel element for a scale η compared to one of interest.

Following (3.15),

$$\begin{aligned} K^{\eta}_{tm,sn} &= \int \mathrm{d}^2\eta r\, P_1(\eta\mathbf{r})\, [b_{tm}(\eta r,\theta)]^*\, b_{sn}(\eta r,\theta) \\ &= \eta^2 \int \mathrm{d}^2\eta r\, P_1(\mathbf{r}) \sum_{t',s'} \left[D_\eta\right]^*_{t,t'} [b_{t'm}(r,\theta)]^* \left[D_\eta\right]_{s,s'} b_{s'm}(r,\theta) \\ &= \eta^2 \left[D_\eta\right]^*_{t,t'} K_{t'm,s'n} \left[D_\eta\right]_{s,s'} \end{aligned}$$

We have used the fact that P_1, as an indicator function, is invariant to shifts of scale, as both the free parameter and the region boundaries shift by the same constant. Using the fact that D_η is real, this means that

$$K_\eta = \eta^2\, D_\eta K D_\eta^T \tag{4.24}$$

is the relation between a kernel at λ_{new} and the one at λ_{old} if we define

$$\eta = \lambda_{\text{old}}/\lambda_{\text{new}} \tag{4.25}$$

The kernels of any apodization problem at different wavelengths are related by a simple transformation.

This relation means that any calculated K, in principle, carries within it the complete information about the geometry of the pupil so far as we are concerned. If we were so fortunate that $[D_\eta]^\dagger = [D_\eta]^{-1}$ then the eigenvalues would be unchanged and the eigenvectors would be a simple rotation. This is not the case; no simple transformation law exists to relate the eigenvalues and eigenvectors at different η. It will be necessary to recalculate these for each K_η, but this is not a computational hardship.

We now turn to derivatives with respect to η. Any $\mathcal{N}$ may be expressed as

$\eta\mathcal{N}_0$ relative to some base $\mathcal{N}_0$. Therefore, $\mathcal{N}\frac{\partial}{\partial\mathcal{N}} = \eta\frac{\partial}{\partial\eta} = \frac{\partial}{\partial \ln\eta}$ is an operator independent of $\mathcal{N}$. This gives us the cleanest results for exploring the behavior of the eigenvectors and eigenfunctions.

$$\begin{aligned}
\eta\frac{\partial}{\partial\eta}D_\eta &= D_\eta\ [\mathbb{I} + r\partial_r] \\
\eta\frac{\partial}{\partial\eta}K &= 4K + [r\partial_r]K + K[r\partial_r]^T \\
\eta\frac{\partial}{\partial\eta}\Lambda_a &= \Lambda_a\left(4 + \langle a|\left([r\partial_r] + [r\partial_r]^T\right)|a\rangle\right) \\
\eta\frac{\partial}{\partial\eta}|a\rangle &= \sum_{\Lambda_b\neq\Lambda_a}\frac{\langle b|\left(\Lambda_a[r\partial_r] + \Lambda_b[r\partial_r]^T\right)|a\rangle}{\Lambda_a - \Lambda_b}|b\rangle
\end{aligned}$$

All K, Λ_a, $|a\rangle$ are evaluated at the chosen η.

We can write these abstract expressions in our chosen (tm) basis, but we must now address the concern of errors introduced by the truncation. Since the operators are independent of m, we only need to examine $t-$indices; since dilation can not mix even t and odd t, we can treat each set separately.

We start with the operator in the eigenvalue derivative $[r\partial_r] + [r\partial_r]^T$, for even t. Using (3.20), we find that for any finite cutoff T this operator has one eigenvalue $(T+2)^2/2 - 4$, and all others are -4. The eigenvector for the unique eigenvalue has entries $\langle t|v_1\rangle = \left(-(-1)^{t/2}\sqrt{\frac{t+1}{T+1}}\right)/(1+T/2)$.

If we rewrite the operator in terms of the eigensystem, we obtain

$$\begin{aligned}\left([r\partial_r] + [r\partial_r]^T\right)_{t\,even} &= \left[\frac{1}{2}(T+2)^2 - 4\right]|v_1\rangle\langle v_1| + [-4]\left(\mathbb{I} - |v_1\rangle\langle v_1|\right) \\ &= -4\mathbb{I} + 2(-1)^{(t+s)/2}\sqrt{(t+1)(s+1)}\end{aligned}$$

independent of the cutoff, with $|v\rangle$ being the eigenvectors. A similar calculation for the odd part gives

$$\left([r\partial_r] + [r\partial_r]^T\right)_{t\,odd} = -4\mathbb{I} - 2(-1)^{(t+s)/2}\sqrt{(t+1)(s+1)}$$

In the eigenvector, $r\partial_r$ and its transpose are triangular matrices, so truncation does not alter the eigensystem but rather only which eigenvalues and eigenvectors are counted. The values of these matrices only grow as $\sqrt{t}$, which is substantially slower than we expect the components $\langle tm|a\rangle$ to decay. Therefore, the effect of truncation on eigenvector and kernel derivatives should be minimal.

These results mean that we can simplify the eigenvalue derivative $\eta\partial_\eta\Lambda_a$ to

$$\eta\frac{\partial}{\partial\eta}\Lambda_a = \Lambda_a \sum_{\substack{t,s \\ \text{t+s even}}} \langle a|t\rangle \left[2\,(-1)^{(t \bmod 2)}(-1)^{(t+s)/2}\sqrt{(t+1)(s+1)}\right]\langle s|a\rangle \tag{4.26}$$

This is of the form $\langle a|\,M\,|a\rangle$ for a symmetric real matrix M, and so can not be negative. We have proven that *all eigenvalues for all apodizations can not decrease with increasing mask number, regardless of pupil geometry.*

4.2.3 Direct check of dilation operator performance

We first check the convergence of the reproduction of in reproducing $b_{t0}(\eta r) = \sum_s [D_\eta]_{t,s} b_{s,0}(r)$. Figure 4.2 compares this for $t = 0$ and $t = 10$ across several η values. Each shows the relative error induced by cutting off the sum at various s. The plots extend up to $r = 15$, which roughly corresponds to the edge of the pupil at the large $\mathcal{N} = 10.0$.

The $t = 0$ mode requires a higher cutoff than the $t = 10$ mode to converge to the same accuracy. This is encouraging, as it implies that an eigenfunction truncated to allow dilation in that mode will automatically allow the other modes to properly dilate, and so the entire Slepian will converge properly using the dilation operator. For accuracy in the field, a relative error of 10^{-5} is desirable so that errors in intensity are of order 10^{-10}, the same as our probable contrast target in the instrument plane.

$\eta < 1$ is faster to converge than $\eta > 1$. This is unsurprising, as matrix entries are polynomial in η, with $(D_\eta)_{t,s}$ containing terms from η^t to η^s; therefore, $\eta > 1$ makes the matrix entries increase without bound as s increases. Convergence, which only occurs for a fixed outermost r, requires precise cancellation. For $\eta > 1$ the dilation matrix places greater emphasis on those higher t, so our truncation of the number of basis modes must always cause an effective upper bound on η.

The $\eta = 1.1$ reproduction of $t = 0$ is within the stated 10^{-5} limit using a cutoff at $t = 14$, within $r \leq 5$ or so. This corresponds to an accurate reproduction inside the pupil up to $\mathcal{N} = 3.2$. Including terms up to $t = 26$ enables reproduction in a pupil up to $\mathcal{N} = 6.5$.

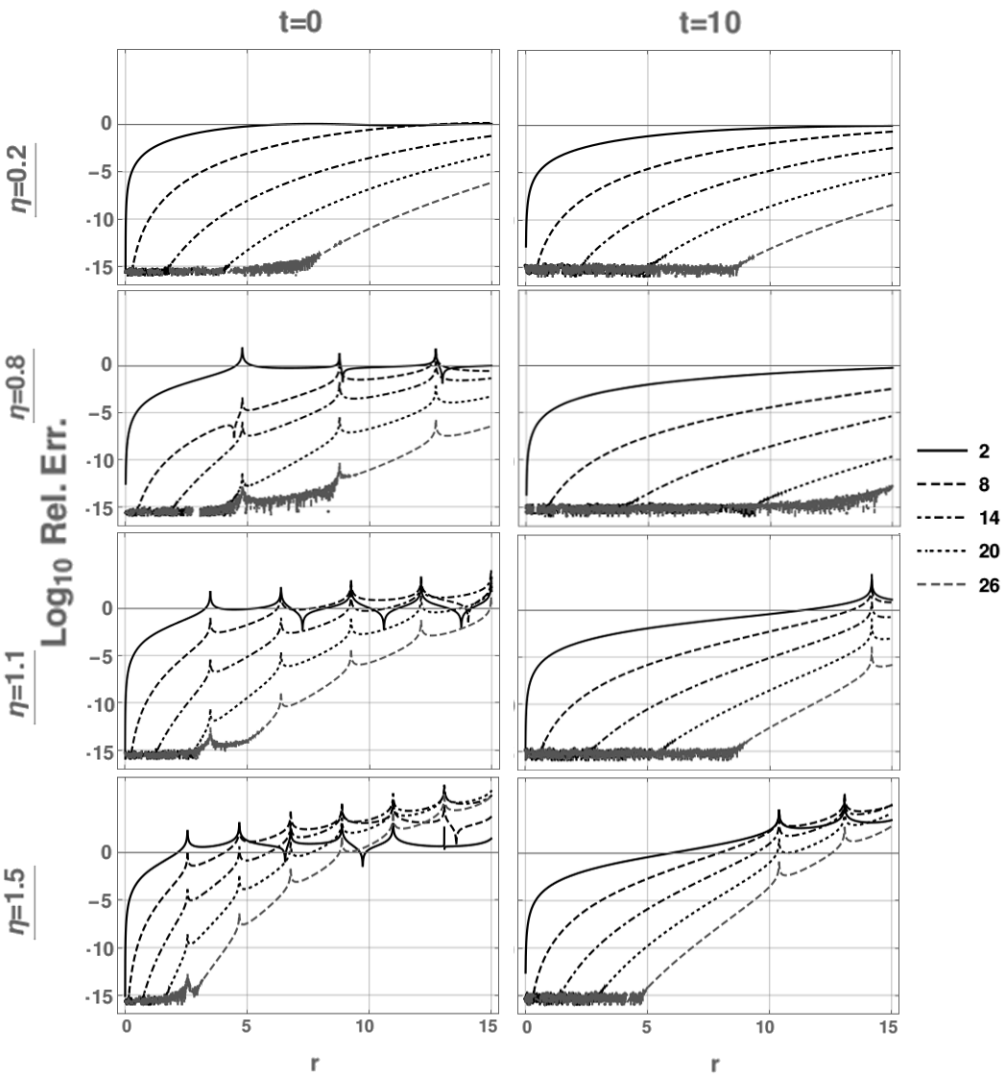


Figure 4.2: Relative errors between $\mathcal{J}_{t+1}(\eta r)$ and $\sum_s [D_\eta]_{t,s} \mathcal{J}_{s+1}(r)$, caused by truncating the sum at different modes. The cutoffs in s are indicated by the legend.

Given these results, as we have stated in 4.2.1, we recommend in practice that $\eta < 1.0$ (i.e. $\lambda_{old} < \lambda_{new}$) be used if we wish to find the Slepian modes for a bandwidth. If necessary to go above this, we suggest $\eta = 1.1$ be considered a cutoff to avoid inclusion of a very large number of extra basis functions.

We now compare eigenvalues from kernels directly calculated by the

integral (3.15) and by the dilation method (4.24). The basis functions were limited to fixed $m = 0$ for a circular geometry, as this gave an acceptable spread of values while allowing for a greatly reduced number of basis functions without compromising results. The coronagraphic parameters were $R_S = 0.2$ for $\mathcal{N} = 2.0, 4.0, 6.0, 8.0, 10.0$. Dilation was from the $\mathcal{N} = 10.0$ kernel.

The same matrix multiplications were carried out to create a "dilated" version of the $\mathcal{N} = 10.0$ kernel, as a control for the error introduced at very high t. The cutoff $T = 30$ in both cases.

Figure 4.3 shows the relative error in the top six eigenvalues from the dilation method, compared to the direct kernel calculation, as well as the actual eigenvalues over that range.

The relative error grew very high, reaching 100% for $a = 6$ by $\mathcal{N} = 2.0$. As this eigenvalue started on the order of 0.1 at $\mathcal{N} = 10.0$ and ended at order 10^{-13}, we do not find this discouraging. Even down to $\mathcal{N} = 6.0$ the relative error remained under 10^{-6} even as Λ_6 itself approached 10^{-6}. The eigenvalues above 10^{-10} at $\mathcal{N} = 2.0$ suffered from relative errors under 10^{-5}, which we judge to be an acceptable change.

We then compared the direct calculation eigenvalues at $T = 40$ to the dilation results from $\mathcal{N} = 10$ to 2 using cutoffs of $T = 30, 20, 14$. Figure 4.4 shows relative error plots for the top six eigenvalues. Errors accumulated more rapidly for lower T. The compromise $T = 20$ had an acceptably low relative error to $T = 40$ (which are presumably very accurate eigenvalues) and a generally acceptable change (three orders of magnitude in the relative error) down to $\mathcal{N} = 6.0$ for the eigenvalues above 10^{-5}.

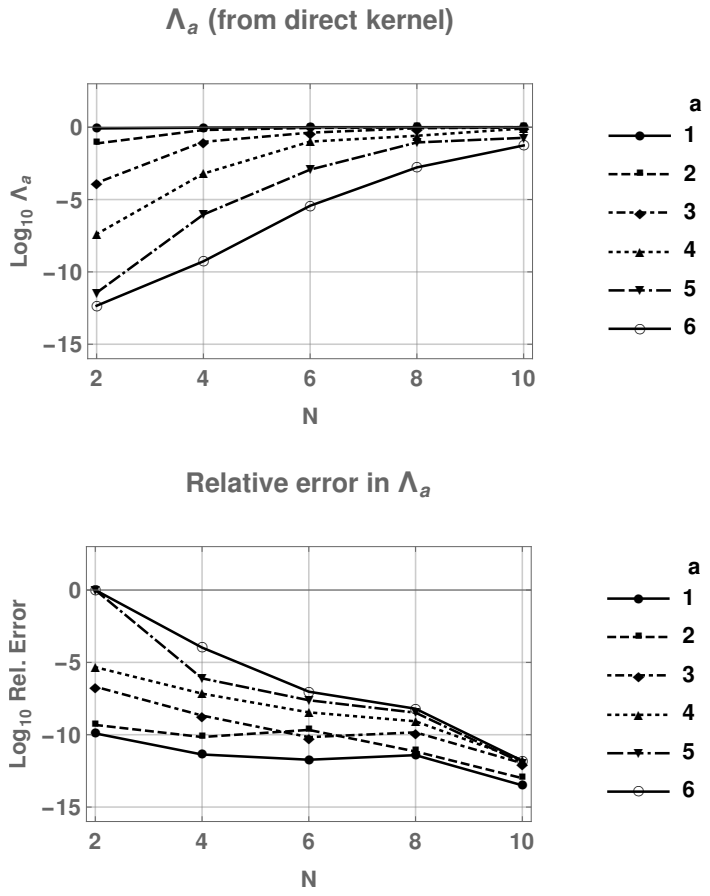
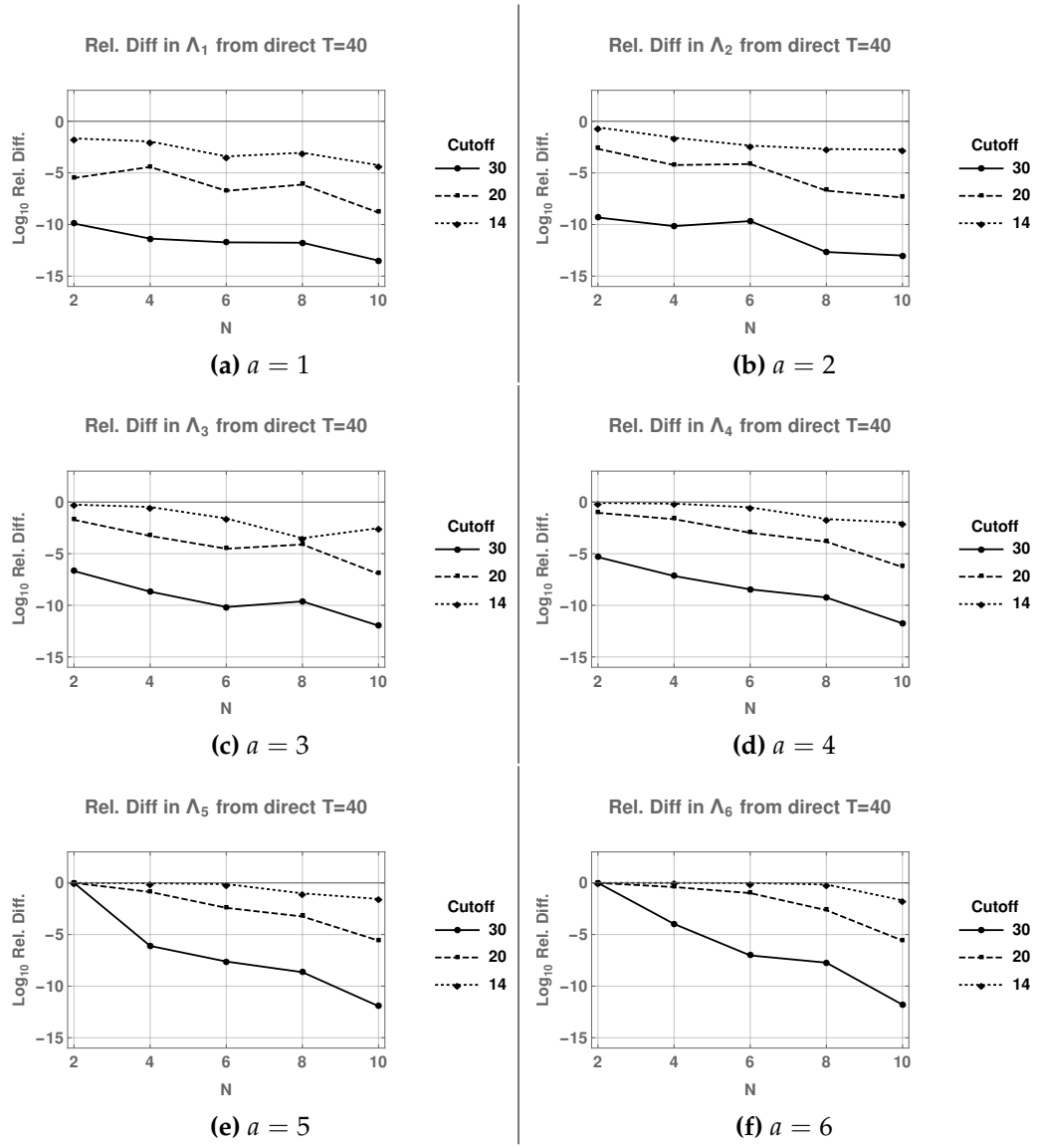


Figure 4.3: The relative error that results from using the kernel dilation method (4.24) instead of a direct calculation of the kernel, and the true eigenvalues for that range of $\mathcal{N}$.



(a) $a = 1$ **(b)** $a = 2$ **(c)** $a = 3$ **(d)** $a = 4$ **(e)** $a = 5$ **(f)** $a = 6$

Figure 4.4: Relative error in the dilated eigenvalues compared to a direct kernel calculation using $T = 40$.

We therefore consider the use of the kernel dilation method to be acceptable over bandwidths up to 25%, and possibly beyond. To use this dilation requires using more basis functions than would be indicated by § 3.3.1 for the lower bandwidths, but is roughly equal to the suggested $2\mathcal{N}$ in § 3.4 for the largest $\mathcal{N}$. Care should be taken with eigenvalues which drop to small values as the precision and truncation errors will accumulate rapidly there.

4.2.4 Broadband propagation

With the formalism established, we can now examine the propagation of light through the coronagraph when the incident wavelength λ_i does not match the design wavelength λ. We will take $\lambda/\lambda_i = \eta_i$, and presume incidence directly onto the pupil apodized by $|\alpha\rangle$.

Before we begin the propagation, we wish to first remark on the possible chromatic aberration Both the pupil apodization and the mask function f will suffer from this effect if they alter the phase of the light, while functions which are positive or zero will not. This must be accounted for by altering the expression for f or the coefficients α_a as appropriate. We have not attempted a rigorous treatment of this, but we will explore a brief generalized possibility.

Let's say that the mask causes phase change, due to an optical element of refractive index n producing some optical path length designed around wavelength λ_0. The phase shift at λ_i will be that at λ multiplied by

$$n(\lambda_i)/n(\lambda_0) \approx 1 + ([\lambda_i/\lambda_0] - 1)\frac{\partial \ln n}{\partial \ln \lambda} + \dots$$

We expect that the second term (and higher order terms) will be sufficiently small, denoted by ϵ, to use a phase-amplitude coupling $e^{ix(1+\epsilon)} \approx e^{ix}(1 + e^{i\pi/2}x\epsilon)$. This means we do not have to address questions of non-integer phase dependence $e^{i\ell(1+\epsilon)\theta}$.

θ itself might be treated as a sawtooth pattern, and expanded as

$$\theta = \pi + e^{i\pi/2} \sum_{n=1,\,odd}^{\infty} \frac{1}{n} \left(e^{in\theta} - e^{-in\theta} \right)$$

We can immediately see the difficulty this poses: the exponentials are *odd* powers $e^{in\theta}$. 3.3.2 showed that this is a difficult case to handle, requiring Taylor expansion of the Struve $H-$function. Given the challenges this poses, we have neglected further study of such chromatic aberration effects. For the remainder, we will approximate that f and the apodization are achromatic, regardless of action on phase.

There are two possible approaches to handling the propagation. The first method is to rewrite the initial apodization in terms of the apodizations of the incident light, which then propagate as in section 4.1. The second is to account for the mismatch in the Fourier transforms between the planes. Both result in the same answer, but emphasize a different method of thinking about the sequence.

For the first method, let's write the coordinates for the design wavelength as (r, θ), so that

$$\phi_a(r, \theta) = \sum_{tm} V_{a,tm} b_{tm}(r, \theta)$$

Using this same r, apodizations which would have been derived for the

incident wavelength are written as

$$\phi'_b(\eta r, \theta) = \sum_{tm} U'_{b,tm} b_{tm}(\eta r, \theta)$$

Using the relation $b_{tm}(r,\theta) = \sum_s [D_{1/\eta}]_{ts} b_{sm}(\eta r, \theta)$ allows us to rewrite ϕ_a as a function of ηr, as

$$\begin{aligned}\phi_a(r,\theta) &= \sum_{sm} \left[\sum_t V_{a,tm} [D_{1/\eta}]_{ts} \right] b_{sm}(\eta r, \theta) \\ &= \sum_b \left(\sum_{sm} \sum_t V_{a,tm} [D_{1/\eta}]_{ts} [U'_{b,sm}]^* \right) \phi'_b(\eta r, \theta) \end{aligned} \tag{4.27}$$

This conversion to a sum over the Slepian modes of the incident wavelength acts no differently from the coefficients α_a.

This approach also requires the scaling of inner products on, and thus Fourier transforms from, $r-$ space. Since we must go from $\int \mathrm{d}^2 r = \eta^{-2} \int \mathrm{d}^2[\eta r]$, the image and Lyot plane field strengths are a factor of η^{-2} compared to (4.12) and (4.15). The instrument plane field must be multiplied by η^{-4}.

While we require the $U'_{b,sm}$, these are just the components of the eigenvector for the kernel dilated by η, (4.24)

$$K_\eta = \eta^2 D_\eta K D_\eta^{\mathrm{T}}$$

and so we can calculate them straightforwardly. A more noticeable problem is the use of $D_{\eta^{-1}}$. We have previously tried to restrict $\eta < 1$, as calculations above $\eta = 1$ were shown to rapidly become unstable. Requiring use of η^{-1} produces bounds on η from below. We defer further comment until after the second approach.

For the second approach in broadband, we convert the incident light into sums over the Slepian modes of the design wavelength. This has the advantage of direct comparisons of all incident waves to a single standard. However, it still requires use of dilation in $1/\eta$.

As in 4.1, the pupil-plane field is just $|\alpha\rangle$, and the power is given by (4.11). First we will re-examine the Fourier transform for an arbitrary basis function,

$$\widehat{b}_{tm}(\rho,\varphi) = \int \mathrm{d}^2 r\, b_{tm}(r,\theta) e^{ir\rho\cos(\theta-\varphi)}$$

If we substitute $r = \eta r'$, $\rho = \rho'/\eta$ then this becomes

$$\begin{aligned}\widehat{b}_{tm}(\rho'/\eta,\varphi) &= \eta^2 \int \mathrm{d}^2 r'\, b_{tm}(\eta r',\theta) e^{ir'\rho'\cos(\theta-\varphi)} \\ &= \sum_s \eta^2 [D_\eta]_{ts} \widehat{b}_{sm}(\rho',\varphi) \\ \therefore \widehat{b}_{tm}(\eta\rho,\varphi) &= \sum_s \eta^{-2} [D_{\eta^{-1}}]_{ts} \widehat{b}_{sm}(\rho,\varphi) \end{aligned} \tag{4.28}$$

in constrast to the previous

$$b_{tm}(\eta r,\theta) = \sum_s [D_\eta]_{ts} b_{sm}(r,\theta)$$

In this approach, we can no longer consider the Fourier transform to be an identity operation when the wavelengths are mismatched. Transforming from object space to image space is an action by $\eta^{-2} D_{\eta^{-1}}$; transforming from image space to object space is an action by D_η. We will refer to these by $\mathfrak{F}_1$ and $\mathfrak{F}_2$, so that the repeated action $\mathfrak{F}_2\mathfrak{F}_1 = \mathfrak{F}_1\mathfrak{F}_2 = \eta^{-2}$.

A single Slepian mode on the pupil therefore transforms to

$$\mathfrak{F}_1 P_1 \left|a\right\rangle = \left[\sum_{tm} \eta^{-2} V_{a,tm} \sum_s [D_{\eta^{-1}}]_{ts} \sum_b (V_{b,sm})^*\right] P_1 \left|b\right\rangle$$

which is to say that we change our coordinates from (tm, sn) to (a, b) as

$$[\mathfrak{F}_1]_{ab} = \sum_{tm,sn} V_{a,tm} \left[\mathfrak{F}_1\right]_{ts} (V_{b,sn})^* \tag{4.29}$$

just as a good operator should.

From all this, the image plane field after the mask is given by

$$\begin{aligned} &[(1 - P_2) + f P_2] \mathfrak{F}_1 P_1 \left|\alpha\right\rangle \\ &= \sum_{a,b} \alpha_a \left[\mathfrak{F}_1\right]_{ab} \sum_c \left[(P_1 - \Lambda_b)\delta_{bc} + \Lambda_b f_{bc}\right] \left|c\right\rangle \end{aligned} \tag{4.30}$$

Effectively at this stage, the mismatch has shifted $\alpha_a \to \sum_b \alpha_b \left[\mathfrak{F}_1\right]_{ba}$. The power is therefore still given by (4.13), with this shift included.

In moving to the Lyot plane, we must act with $\mathfrak{F}_2$. However, this is a linear operator; and, as they are sums over $\left|a\right\rangle$, they commute with the projection operators P_1, P_2 and the mask operator f. For the Lyot plane, then, the two Fourier transforms cancel (as is well-known) and we are left with the same

field as in (4.14), up to the powers of η.

$$\sum_a \eta^{-2}\alpha_a \left[P_3 \left(P_1 - \Lambda_a \right) |a\rangle + \Lambda_a \sum_b \left(f_{ab} \right) P_3 |b\rangle \right]$$

$$\sum_a \eta^{-2}\alpha_a \left[\sum_b ([1 - \Lambda_a]\, \delta_{ab} + \Lambda_a f_{ab}) P_3 P_1 |b\rangle + \sum_b \Lambda_a \left(f_{ab} - \delta_{ab} \right) P_{3+} |b\rangle \right]$$

$$\sum_a \eta^{-2}\alpha_a \left[\sum_b \left([1 - \Lambda_a]\delta_{ab} + \Lambda_a f_{ab} \right) P_3 |b\rangle - P_{3+} |a\rangle \right] \tag{4.31}$$

The α_a are the original, unaltered coefficients from the pupil plane.

The instrument plane fields follow directly by acting on this with $\mathfrak{F}_1 = \eta^{-2}[D_{1/\eta}]$ if we wish to express them as a sum over the Φ_D from the design modes. This will give us the spread of the light in units of λ/D_P. This is exactly the same as the first method, as two calculations of the same physical result should be. The only difference is that the first method produces the field in units of λ_i/D_P, which we could also achieve here by appropriate variable redefinitions.

From both of these methods, we have found that the instrument plane fields for the incident wavelength can be found from those of the design wavelength by acting on the Φ_D from that λ with the operator $\eta^{-4}[D_{\eta^{-1}}]$, for $\eta = \lambda/\lambda_i$. The practical limitations of the dilation operator means that this could be a useful strategy for $0.9 \lesssim \eta \lesssim 1.1$, which matches the 20% bandwidth desired (Krist et al., 2011) for Earth-twin analysis. For wavelengths outside of this band, either direct numerical integration should be used, or we must accept the use of an increased radial mode cutoff T beyond that originally suggested for the kernel dilation method.

Alternatively, we can directly calculate two kernels, each anchoring one end of the bandwidth. The large-wavelength kernel can be used with the dilation method to produce intermediate kernels for eigenvalues and eigenfunctions. The small-wavelength kernel can be used for the coronagraphic propagation. This will also provide an error estimate for the broadband results.

In the event that we have a source which emits light at an intensity per unit wavelength $I(\lambda_i)$, we can find the composite action in the instrument plane by acting with

$$\lambda \int \mathrm{d}\eta^{-1}\; I(\lambda\eta^{-1})\,[\eta^{-1}]^4\,[D_{\eta^{-1}}] \tag{4.32}$$

on the response Φ_D of the coronagraph to the design wavelength. Since all entries $[D_\eta]_{t,s}$ are polynomials from order t to order s, each entry in this combined operator will be a relatively simple integral depending on the model $I(\lambda_i)$.

This is in the achromatic-f and α_a approximation. If we wish to include those effects, then the integral will include those terms, and we no longer have this simple action on Φ_D. Instead, Φ_D must be split and f and α included in the integral. In such a case it may be easiest to calculate transmission of the basis functions b_{tm} as a sum of the initial Slepian modes, and then recombine the final results with the chromatic dispersion effects of the mask and pupil apodization.

4.3 Perturbations and off-axis plane waves

Until this point, we have assumed that our light is a constant plane wave arriving perpendicularly to the pupil. This is obviously an idealization. Turbulence or deviations of the mirror/lens from the ideal shape introduce non-constant wavefronts. Stars are not point objects; planets are off-axis, as are the stars themselves for misalignment errors.

Study of these deviations to date have mostly relied on numerical simulations (Laurent et al., 2018) (Ruane et al., 2018), though some analytical progress has been made (Leboulleux et al., 2018). We develop here the behavior of our formalism under various simple kinds of non-planar wavefronts. The function $g(r, \theta)$ will be our symbol for such a wavefront. Perturbative effects therefore arrive at the pupil plane as $(1 + g)$. We will argue that we can in principle treat *any* wavefront as a linear operator in our system. Our practical limits in mode number T will restrict the functions for which this is a good approach.

4.3.1 Pupil-limited operators as linear operators

Before looking at specific values, we must first consider the validity of handling our g as linear operators, as we have been attempting to do throughout this chapter. We have already shown that some functions (e.g. r^k) can be defined tolerably well in general, despite the fact that the integral $\int_0^\infty \mathrm{d}r\, r^k J_{t+1}(r)$ does not exist.

We circumvented this requirement § 3.3.3 by using recursion relations to determine the exact coefficients, and only required convergence inside the

pupil. (More properly, inside the circumscribing circle where $r < \mathcal{N}\pi/2$.) This leads us to look into defining arbitrary functions inside the pupil.

Recall that the eigenvectors $|a'\rangle$ corresponding to $K' = P_1P_2P_1$, the forwards kernel, are entirely contained inside P_1 and form a complete basis for all finite and square-integrable functions on that space. We have previously pointed out 2.5 how we can write $P_1 = \sum_a' |a'\rangle \langle a'| = \sum_a \Lambda_a^{-1} P_1 |a\rangle \langle a| P_1$. Even though the eigenvalues are exponentially decreasing, this may be well-defined if the relevant functions $|tm\rangle$ in each $|a\rangle$ die even faster.

Using our identity for P_1, we have that

$$\begin{aligned} P_1 g P_1 &= \sum_{a,b} P_1 \frac{|a\rangle \langle a|}{\Lambda_a} P_1 g P_1 \frac{|b\rangle \langle b|}{\Lambda_b} P_1 \\ &= \sum P_1 |j\rangle \langle k| P_1 \cdot L_{j,i} \widetilde{g}_{i,h} L_{h,k} \end{aligned} \tag{4.33}$$

where

$$\widetilde{g}_{j,k} = \int \mathrm{d}^2 r \, \Omega_1(\mathbf{r}) g(\mathbf{r}) b_j^*(\mathbf{r}) b_k(\mathbf{r}) \tag{4.34}$$

are the components of the pupil limitation of the operator in our chosen basis, and

$$\begin{aligned} L &= \sum_a \frac{|a\rangle \langle a|}{\Lambda_a} \\ L_{j,k} &= \sum_a \frac{V_{a,j} V_{a,k}^*}{\Lambda_a} \end{aligned} \tag{4.35}$$

We can immediately see the danger involved in this, as the very small values of Λ_a will eventually include far too much error. We will need to truncate the sum in a when $1/\Lambda_a$ ceases to be reliable, at the very latest.

Let's examine the behavior of such an operator on a Slepian mode. We have that

$$\begin{aligned} P_1 g P_1 \left|a\right\rangle &= \sum_{ijk} P_1 \left|i\right\rangle \ \left\langle j\right| P_1 \ \cdot (L\widetilde{g}L)_{ij} V_{a,k} \left|k\right\rangle \\ &= \sum_{ijk} P_1 \left|i\right\rangle \ \cdot (L\widetilde{g}L)_{ij} K_{jk} V_{a,k} \\ &= \sum_{ij} P_1 \left|i\right\rangle \ (L\widetilde{g}L)_{ij} \Lambda_a V_{a,j} \end{aligned}$$

From our definition of L_{ij}, we have that $\sum_j (L\widetilde{g}L)_{ij} \Lambda_a V_{aj} = \sum_{km} L_{ik} \widetilde{g}_{km} V_{a,m}$. Alternatively, we could say that $\sum_j (L\widetilde{g}L)_{ij} K_{jk} = \sum_j L_{ij} \widetilde{g}_{jk}$. Either way, our action is now

$$\begin{aligned} P_1 g P_1 \left|a\right\rangle &= \sum_i (L\widetilde{g}V_a)_i P_1 \left|i\right\rangle \\ &= \sum_{ib} (L\widetilde{g}V_a)_i (V_{b,i})^* \ P_1 \left|b\right\rangle \end{aligned}$$

i.e.

$$P_1 g P_1 \left|a\right\rangle = \sum_b \left[\sum_{jk} \frac{(V_{b,j})^*}{\Lambda_b} \widetilde{g}_{jk} V_{a,k} \right] P_1 \left|b\right\rangle \tag{4.36}$$

For convenience, we may define

$$\Gamma_{ab} \equiv \sum_{jk} \frac{(V_{b,j})^*}{\Lambda_b} \widetilde{g}_{jk} V_{a,k} \tag{4.37}$$

in contrast to

$$\widetilde{g}_{ab} \equiv \sum_{jk} (V_{b,j})^* \widetilde{g}_{jk} V_{a,k}$$

Errors in calculation of Λ_b therefore produce absolute errors in the operator of order $\frac{\delta \Lambda_b}{\Lambda_b}$, the relative error in the eigenvalue.

This Γ matrix is a bit unusual; it obeys the following properties.

$$\Gamma_{ab} \times \Lambda_b = \widetilde{g}_{ab}$$

$$\left(\frac{\Lambda_b}{\Lambda_a}\right) \times \Gamma_{ab} = (\Gamma_{ba})^*$$

We demonstrate the reproduction of the constant pupil for our standard example. To do so, we calculate the coefficients to recreate $P_1 |1\rangle$, the constant function in the pupil. This is fairly simple,

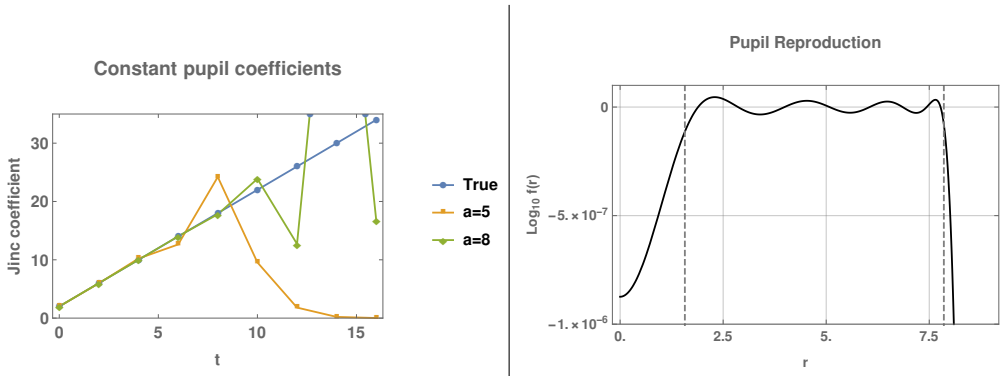
$$\begin{aligned} P_1 |1\rangle &= \sum_a P_1 |a\rangle \frac{\langle a| P_1 |1\rangle}{\Lambda_a} \\ &= \sum_{tm} \left[\sum_{sn,a} \frac{(V_{a,sn})^* \langle sn| P_1 |1\rangle V_{a,tm}}{\Lambda_a} \right] P_1 |tm\rangle \end{aligned}$$

The inner product $\langle sn| P_1 |1\rangle V_{a,tm}$ is just the integral over the pupil of $[b_{sn}(r,\theta)]^*$. This demonstration is greatly aided by the fact that

$$1 = \sum_{\substack{t=0 \\ \text{t even}}}^{\infty} 2(t+1)\mathcal{J}_{t+1}(r)$$

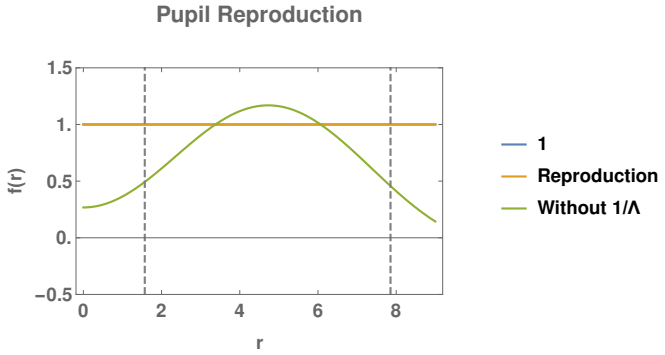
so that we can compare the calculated coefficients to the true ones, and only require $m = 0$ in this example. Figure 4.5 compares the resulting coefficients and the relative error; the accuracy is good to within 10^{-7} for this example.

We also show, in figure 4.6, the result of the reproduction compared with the case where we do not include the $1/\Lambda$ factors. This is the projection of the pupil onto the space of mask-limited functions, $P_2 P_1 |1\rangle$, when displayed in $r-$space.



(a) (b)

Figure 4.5: Reproduction of the constant pupil for the standard example. Left, comparison of the calculated coefficients to the true ones including up to $a = 5$ and $a = 8$ in the sum. Right, log relative error in the $a = 8$ approximation. The divergence at the outside and inside radii is apparent.



(a)

Figure 4.6: Reproduction of the constant pupil for the standard example. The reproduction overlays the constant within this view. We show the case where we do not include $1/\Lambda$ factors, which is the projection of the pupil onto the space of mask-limited functions.

4.3.2 Polynomial perturbations

Effects for which we can effectively decompose g as a polynomial sum $\sum_{k,\ell} c_{k,\ell} r^k e^{i\ell\theta}$ follow immediately from the development of our simple operators in 3.3.3,

equation (3.18).

$$r^k e^{i\ell\theta}\,|t,m\rangle = \sum_{p=0}^{\infty} 2^k\sqrt{t+1}\left[\frac{(-1)^p\sqrt{t+k+2p+1}(k+p-1)!(t+k+p)!}{p!(k-1)!(t+p+1)!}\right]$$
$$\times\,|t+k+2p,m+\ell\rangle$$

$$\left[r^k e^{i\ell\theta}\right]_{sn,tm} = 0 \text{ if } (t+k-s) \text{ is not an even non-negative integer}$$
$$= 2^k\,(-1)^{(s-t-k)/2}\sqrt{t+1}\sqrt{s+1}\left[\frac{\left(\frac{s-t+k-2}{2}\right)!\left(\frac{s+t+k}{2}\right)!}{\left(\frac{s-t-k}{2}\right)!(k-1)!\left(\frac{s+t-k+2}{2}\right)!}\right]$$
$$\times\,\delta_{m+\ell,n} \tag{4.38}$$

assuming that $|\ell| \leq k$ and $(k-|\ell|)/2$ is an integer. We can write this as the effect on the Slepian modes using the usual transformation matrix (eigenvector matrix)

$$\left[r^k e^{i\ell\theta}\right]_{ab} = \sum_{tm,sn} V_{a,sn}\left[r^k e^{i\ell\theta}\right]_{sn,tm}(V_{b,tm})^*$$

This style of perturbation directly applies for Seidel or primary wavefront errors, which can be written using the Zernike polynomials in $r_1/R_P = \frac{r}{\mathcal{N}\pi/2} \equiv x$ ((Born and Wolf, 1999), table 9.2, with some modification):

While we have already demonstrated the convergence of the r^k matrix approach in 3.3.3, it is somewhat informative to contrast it with the $1/\Lambda$ approach developed just above, in 4.3.1. Figure 4.7 shows the relative errors in the reproduction of $(r/\mathcal{N}\pi/2)^2$ and $(r/\mathcal{N}\pi/2)^8$ acting on the first and fifth Slepian mode in the standard example geometry. Cutoffs in a for the $1/\Lambda_a$ were $a = 8$ for r^2 and $a = 6$, 10 for r^8 acting on ϕ_1 and ϕ_5.

Name	Function	Expansion
Tip/Tilt	$R_1^1(x)e^{\pm i\theta}$	$xe^{\pm i\theta}$
Field Curvature	$R_2^0(x)$	$2x^2-1$
Astigmatism	$R_2^2(x)e^{\pm 2i\theta}$	$x^2e^{\pm 2i\theta}$
Coma	$R_3^1(x)e^{\pm i\theta}$	$(3x^3-2x)e^{\pm i\theta}$
Spherical Aberration	$R_4^0(x)$	$6x^4-6x^2+1$

Table 4.2: Primary Seidel aberrations, with $x = r_1/R_P = \frac{2}{\mathcal{N}\pi}r$. Table adopted from (Born and Wolf, 1999), table 9.2.

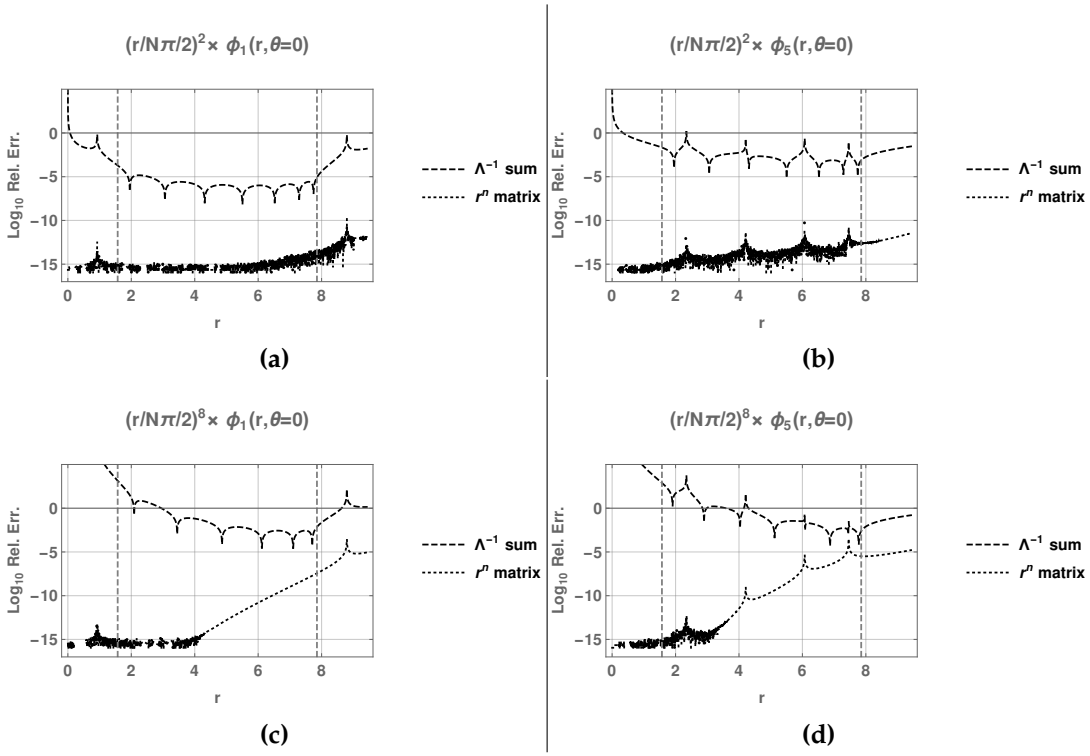


Figure 4.7: Comparison of the relative errors produced in reproducing various $r^n\phi_a$ with the matrix derived from recursion relations (3.18) and the $1/\Lambda$ method (4.33).

We can see that the $1/\Lambda$ reconstruction does considerably worse, though the values of the field there are themselves extraordinarily small. Further examination of the $r^8\phi_1$ case shows that this result comes about from errors

in Λ, as expected. If only those a are included for which $\Lambda_a > 10^{-10}$, then the (normalized) coefficients generated for the sum over Slepian modes are within a relative 10^{-4} of each other.

When written as a sum over the basis functions, this truncation causes errors in the coefficients of the lowest t modes, which are therefore not fully canceled as the true expansion of $r^k \mathcal{J}_{t+1}(r)$ requires. We therefore anticipate that functions which do not require this precise cancellation (i.e. begin at r^0 or r^1 when expanded about $r = 0$) will be more amenable to the $1/\Lambda$ expansion than these particular cases.

4.3.3 Off-axis plane waves

So far, we have looked at cases where the coronagraph is pointed directly at a perfect point source. In reality, stars have finite size and telescopes inevitably carry some pointing error. These are small angular deviations. If there truly is an off-axis planet that we are attempting to image — the purpose of the entire coronagraph! — then it, too, will not illuminate the pupil uniformly. We must therefore study the effect of such occurrences.

We will assume that the incident light is a plane wave tilted by an angle χ relative to the optical axis. ($\chi = 0$ is our old ideal case.) The phase over a flat pupil will therefore be

$$e^{i\mathbf{k}_i \cdot \mathbf{r}_1} = e^{i \sin\chi \frac{2\pi r_1}{\lambda_i} \cos(\theta - \theta_i)}$$

with θ_i the angle the wavevector makes in the pupil plane. Assuming $\chi \ll 1$,

some math shows that this is equal to

$$\begin{aligned} e^{i\mathbf{k}_i\cdot\mathbf{r}_1} &= e^{i\eta\frac{2}{\mathcal{N}}\frac{\chi}{\lambda/D_P}\cdot r\cos(\theta-\theta_i)} \\ &= e^{i\left(\frac{\eta\chi}{R_{M/L}}\right)\cdot r\cos(\theta-\theta_i)} \\ &= e^{i\pi\left(\frac{\eta\chi}{\lambda/D_P}\right)\cdot\left(\frac{r}{\mathcal{N}\pi/2}\right)\cos(\theta-\theta_i)} \end{aligned} \tag{4.39}$$

with $\eta = \lambda/\lambda_i$ the ratio of the design wavelength and the incident wavelength.

The phase variation over the entire pupil is $2\pi\cdot[\eta\chi/(\lambda/D_P)] \equiv 2\pi\omega$. Empirically, we have found that for $\omega \leq 0.1$ we only need to expand the exponential to r^2 to handle the effect to within our usually desired 10^{-5} relative tolerance for field values. This should be sufficient for many small problems at our design specifications.

For larger values of ω, we can not use this direct Taylor expansion. Neither can we attempt the same trick that we did with the polynomials, where we used recursion identities to define a universally valid matrix in a usable form. While the integrals $\int\ \mathrm{d}^2r\,[b_{tm}(r,\theta)]^*[b_{sn}(r,\theta)]e^{i\omega r/(\mathcal{N}\pi/2)\cos(\theta-\theta_i}$ are convergent for $\omega < \mathcal{N}$, this only means that some portion of the function can be represented in this manner. The phase shift simply moves what would have been the on-mask pattern off-mask, where our basis functions cannot by themselves reproduce it.

We therefore turn to attempting the pupil-limited $1/\Lambda_a$ approach of (4.33) through (4.35). As stated before, we must be careful as we must truncate in a before the error in the calculated $1/\Lambda_a$ becomes overwhelming. We can

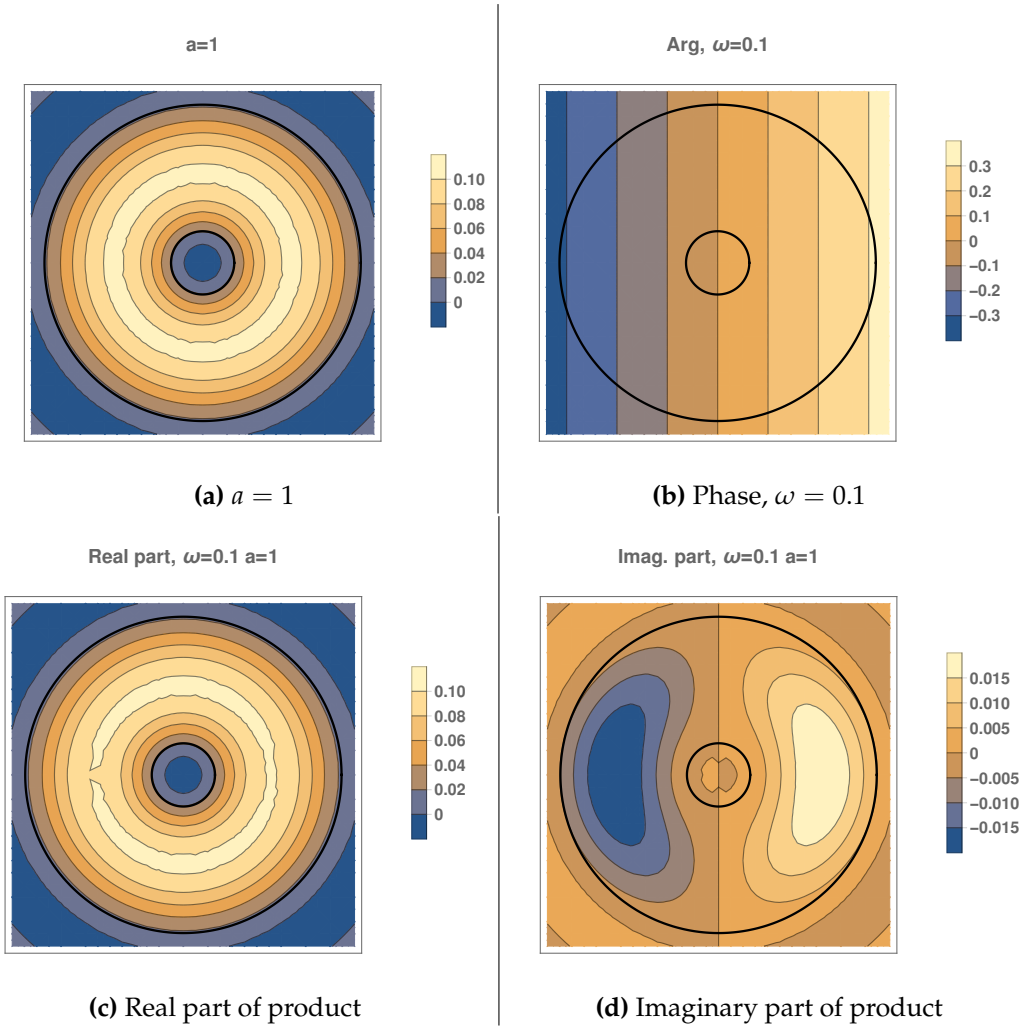
anticipate that this will require a very large number of basis functions, however, using the Jacobi-Anger expansion (*NIST Digital Library of Mathematical Functions*) 10.12.3

$$e^{ix\cos(\theta-\theta_i)} = \sum_{m=-\infty}^{\infty} J_m(x)e^{im(\theta-\theta_i+\pi/2)}$$

This indicates that our reproduction is likely similar to the Bessel function algebra (3.19) for convergence.

As a test, we have run the standard test geometry (circular pupil, $R_S = 0.2$, $\mathcal{N} = 5.0$) with incident plane waves at $\omega = 0.1$, 1.0, and 5.0; $\theta_i = 0$ for simplicity. The basis functions used a maximum $T = 14$, but cut off maximum in m at 10 (limiting us to 108 basis functions). The sum over a used all $\Lambda_a > 10^{-6}$; this was the first 63 of the modes.

Figures 4.8 – 4.13 show the original apodization, the argument of the incident plane wave, the real and imaginary parts of the product, and the log relative error in the real and imaginary parts of the reconstruction for each of the three ω under consideration.



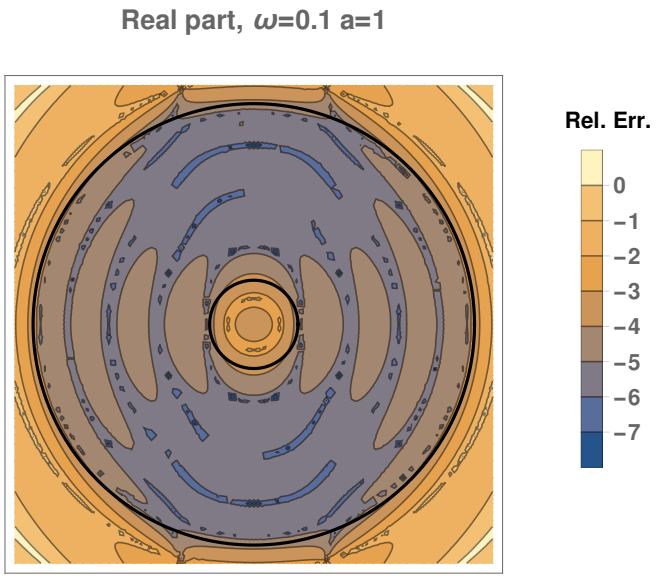
(a) $a = 1$

(b) Phase, $\omega = 0.1$

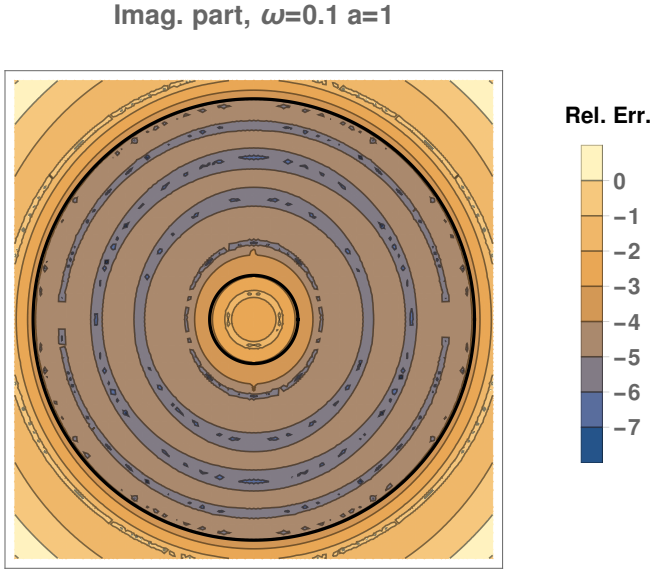
(c) Real part of product

(d) Imaginary part of product

Figure 4.8: Original apodization, incident wave phase, and real and imaginary parts of the resulting product for $\chi/(\lambda/D_P) = \omega = 0.1$. Black lines indicate the inner and outer radii of the pupil.

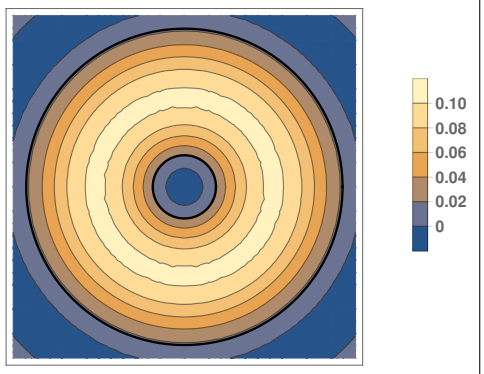


(a) Relative error, real part

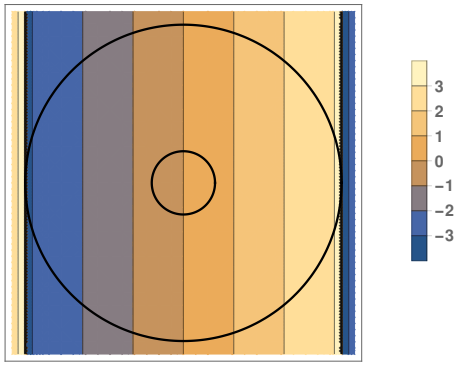


(b) Relative error, imaginary part

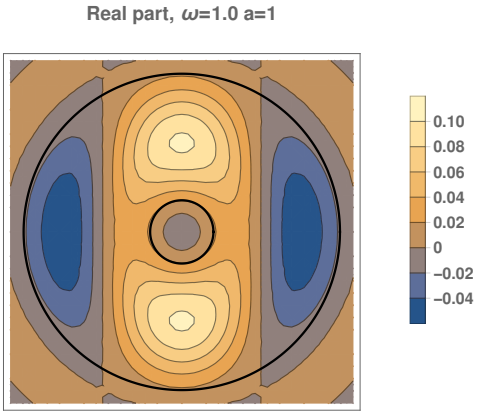
Figure 4.9: $\log_{10}$ Relative error in reconstructing the off-axis apodized light from figure 4.8. Black lines indicate the inner and outer radii of the pupil.



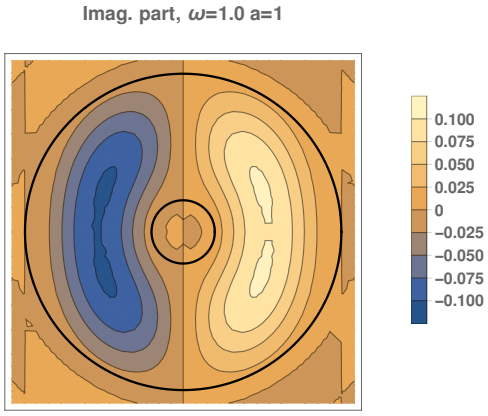
(a) $a = 1$



(b) Phase, $\omega = 1.0$

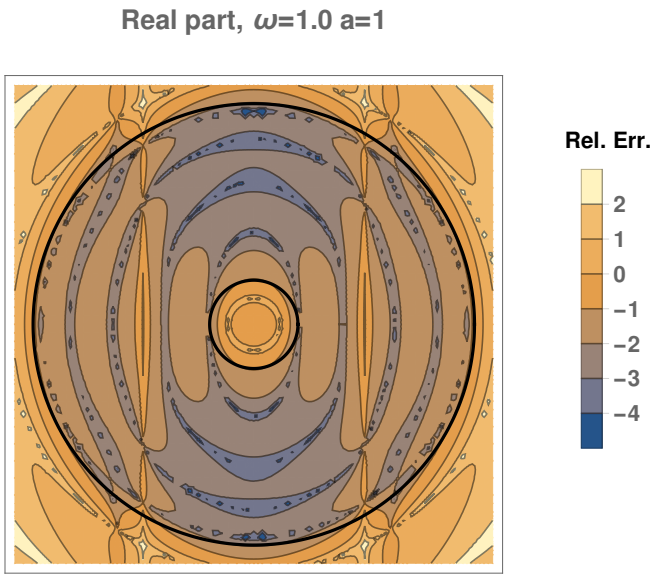


(c) Real part of product

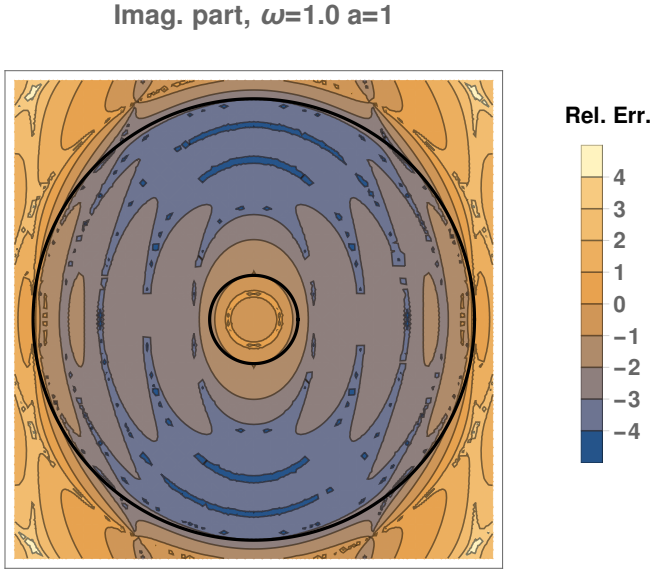


(d) Imaginary part of product

Figure 4.10: Original apodization, incident wave phase, and real and imaginary parts of the resulting product for $\chi/(\lambda/D_P) = \omega = 1.0$. Black lines indicate the inner and outer radii of the pupil.

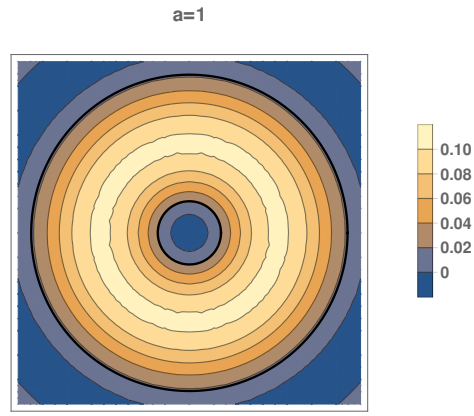


(a) Relative error, real part

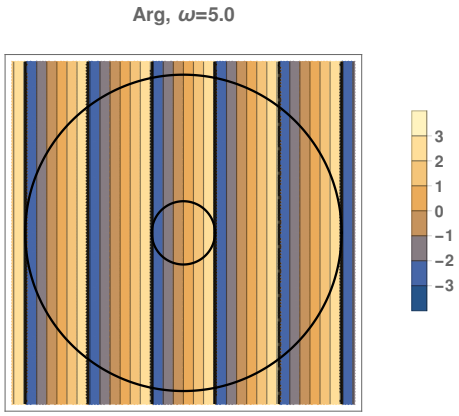


(b) Relative error, imaginary part

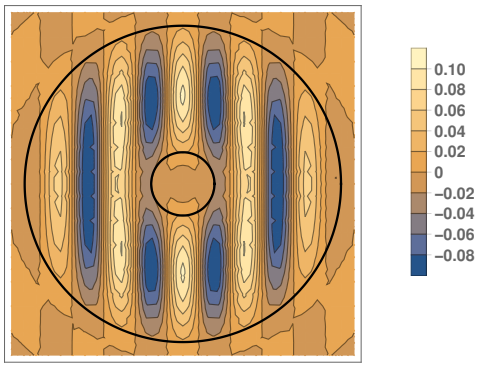
Figure 4.11: $\log_{10}$ Relative error in reconstructing the off-axis apodized light from figure 4.10. Black lines indicate the inner and outer radii of the pupil.



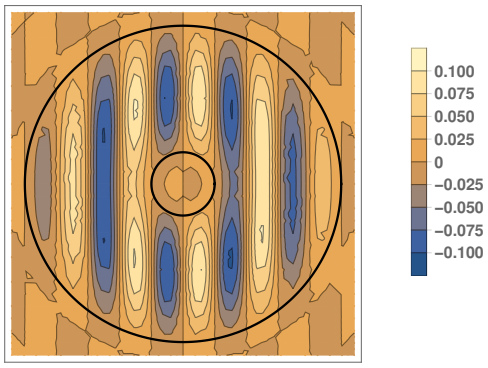
(a) $a = 1$



(b) Phase, $\omega = 5.0$

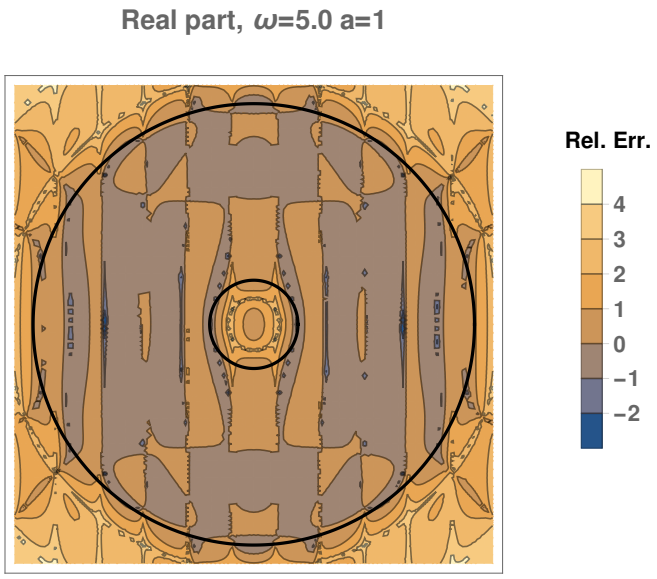


(c) Real part of product

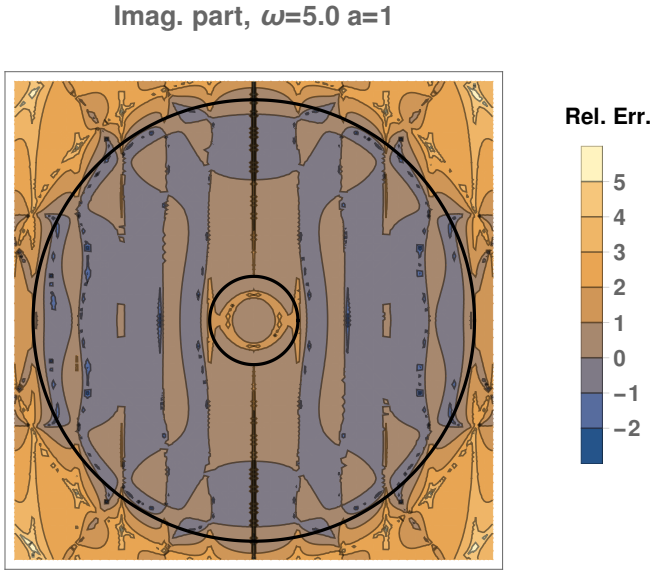


(d) Imaginary part of product

Figure 4.12: Original apodization, incident wave phase, and real and imaginary parts of the resulting product for $\chi/(\lambda/D_P) = \omega = 5.0$. Black lines indicate the inner and outer radii of the pupil.



(a) Relative error, real part



(b) Relative error, imaginary part

Figure 4.13: $\log_{10}$ Relative error in reconstructing the off-axis apodized light from figure 4.12. Black lines indicate the inner and outer radii of the pupil.

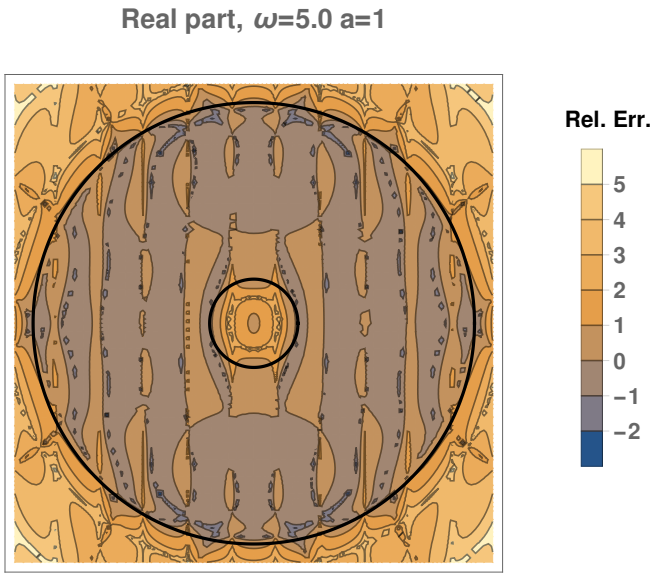
We can see that we do not meet the 10^{-5} target for any of the three off-axis examples with our basis functions up to $t_{max} = 14$, $m_{max} = 10$ and included $1/\Lambda_a$ for $\Lambda_a > 10^{-}6$. $\omega = 0.1$ is nonetheless reasonably well reconstructed, with relative errors mostly below 10^{-4}. $\omega = 1.0$ sees errors from 10^{-4} to 10^{-2}, with stronger errors along the zeros of the real part. For $\omega = 5.0$, a source which is one mask diameter off-axis, the reconstructed relative errors reach from $\approx 10\%$ to $\approx 200\%$.

The relative error from a second attempt at the $\omega = 5.0$, using $t_{max} = 16$, $m_{max} = 16$ (153 basis functions), and all eigenvalues above 10^{-9}, is shown in 4.14. Minor improvement is visible, but the size of the relative errors still goes above 100%. We expect that the most significant improvement would occur at $m = 20$, as that is the next even multiple of ω.

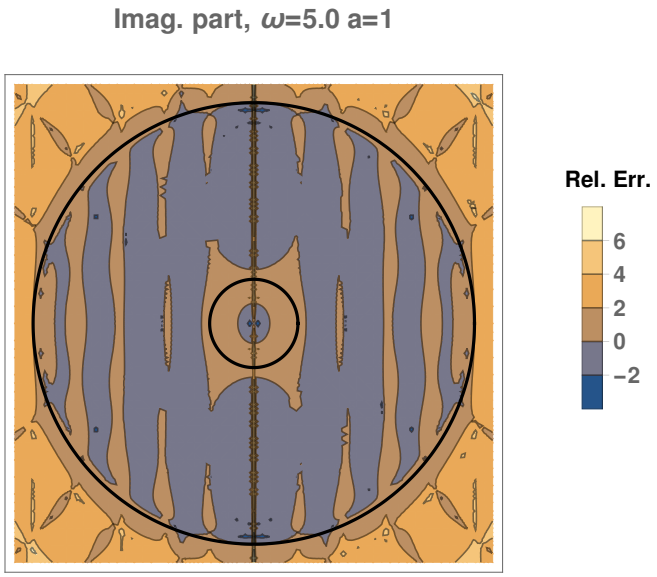
We are forced to conclude that reproducing the pupil-plane fields in this formalism (or the image plane fields) for strongly off-axis sources requires a tremendous number of basis functions, to an angular mode of a few ω and including very small Λ_a.

Following the discussion at the beginning of § 4.1, we are still able to produce the fields in the Lyot plane. The in-pupil region will come from summing mask-limited pattern and the original pupil-field function, which can be calculated simply by the multiplication of the apodization and the required phase.

In the instrument plane, we have a means of underestimating the light from our off-axis source. Recall that the directly on-axis field on the mask contains the fraction $\sum_a |\alpha_a|^2 \Lambda_a$ of the light that passes through our (possibly



(a) Relative error, real part



(b) Relative error, imaginary part

Figure 4.14: Relative error in reconstructing the off-axis apodized light from figure 4.8, but using a cutoff in t and m at 16, and including all eigenvalues $\Lambda_a \geq 10^{-9}$. Black lines indicate the inner and outer radii of the pupil.

apodized) pupil. A simplistic estimate would be to take this pattern and shift it in the instrument plane by ω. For $\omega < \mathcal{N}$ there will be a portion of this off-axis light which fell on the mask, and so is altered; this portion is in the g_{ab} matrix, and acts without the L matrices when propagating.

This approximation neglects the differences from the Lyot stop and the pupil. As this light is comparatively bright in the Lyot stop, neglecting it is likely to lead to major differences. This stopgap approach would therefore be a strong underestimate of the brightness of the off-axis source in the instrument plane, possibly by an order of magnitude. Only in the event that the Lyot stop does not contain areas which are blank in the pupil plane ($P_{3+} = 0$) might this approximation be useful.

4.3.4 Propagation of perturbations and off-axis plane waves

We will now trace the progress of an incident wave g. It is more useful to write the incident field explicitly as a direct plane wave and a perturbation, $1 + \epsilon P_1 g P_1$. We have introduced the parameter ϵ, which is intended to be a small value suitable for order-by-order expansion to show relative importance of different terms.

We will also give perturbations to the residual and throughput. As these expressions are fairly complex, we will use a shorthand notation to indicate the coefficients of various powers in ϵ. The actual coefficients can be inferred from the relevant expressions for the power.

With this new form, the pupil plane field is just

$$\sum_a \alpha_a \left[\delta_{ab} + \epsilon\Gamma_{ab}\right] P_1 \left|b\right\rangle$$

and has power

$$\begin{aligned}
&\sum_a |\alpha_a|^2 \Lambda_a + 2\epsilon \,\mathrm{Re} \sum_{ac} \alpha_a^* \widetilde{g}_{ab} \alpha_c + \epsilon^2 \sum_{abc} \alpha_a^* \Gamma_{ab}^* \Lambda_b \Gamma_{cb} \alpha_c \\
&= \sum_a |\alpha_a|^2 \Lambda_a + \epsilon \sum_{ac} \alpha_a^* \left[\widetilde{g}_{ac} + (\widetilde{g}_{ac})^*\right] \alpha_c + \epsilon^2 \sum_{abc} \alpha_a^* \left[(\widetilde{g}_{ab})^* \frac{1}{\Lambda_b} \widetilde{g}_{cb}\right] \alpha_c
\end{aligned} \tag{4.40}$$

The image plane field is also straightforward,

$$\sum_{abc} \alpha_a \left[(P_1 - \Lambda_b)\delta_{bc} + \Lambda_b f_{bc}\right] \left[\delta_{ab} + \epsilon\Gamma_{ab}\right] \left|c\right\rangle$$

though we neglect to write the power. We can almost, but not quite, summarize this by linear shifts in the coefficients; we are held back by the operator P_1.

The Lyot plane field is

$$\begin{aligned}
&\sum_{abc} \alpha_a (\delta_{ab} + \epsilon\Gamma_{ab})(\delta_{bc} + \Lambda_b[f_{bc} - \delta_{bc}]) P_3 P_1 \left|c\right\rangle \\
&\qquad + \sum_{abc} \alpha_a (\delta_{ab} + \epsilon\Gamma_{ab})(\Lambda_b[f_{bc} - \delta_{bc}]) P_{3+} \left|c\right\rangle
\end{aligned} \tag{4.41}$$

when split into the different regions. The power is

$$\left(\sum_{all}\left[\alpha_a^*(\delta_{ab}+\epsilon\Gamma_{ab}^*)(\delta_{bc}+\Lambda_b[f_{bc}^*-\delta_{bc}])\right]\left[\alpha_x(\delta_{xy}+\epsilon\Gamma_{xy})(\delta_{yz}+\Lambda_y[f_{yz}-\delta_{yz}])\right]\right.$$
$$\left.\cdot\langle c|\,P_3P_1\,|z\rangle\right)$$
$$+\left(\sum_{all}\left[\alpha_a^*(\delta_{ab}+\epsilon\Gamma_{ab}^*)(\Lambda_b[f_{bc}^*-\delta_{bc}])\right]\cdot\left[\alpha_x(\delta_{xy}+\epsilon\Gamma_{xy})(\Lambda_y[f_{yz}-\delta_{yz}])\right]\right.$$
$$\left.\cdot\langle c|\,P_{3+}\,|z\rangle\right) \tag{4.42}$$

We can rewrite this by replacing $P_3P_1 \rightarrow P_1 - P_{1+}$. The $\langle c|\,P_{1+}\,|z\rangle$ and $\langle c|\,P_{3+}\,|z\rangle$ terms follow immediately, and we forgo rewriting them, but the resulting $\langle c|\,P_1\,|z\rangle = \Lambda_c\delta_{cz}$ term (with all summation assumed) expands to

$$\left[|\alpha_c|^2\Lambda_c(1-\Lambda_c)^2+\Lambda_c(1-\Lambda_c)\Lambda_y\left[\alpha_c^*f_{yc}\alpha_y+\alpha_y^*f_{yc}^*\alpha_c\right]\right.$$
$$\left.+\alpha_b^*\alpha_y\Lambda_b\Lambda_c\Lambda_yf_{bc}^*f_{yc}\right]$$
$$+\epsilon\left[\alpha_a^*\left(\Lambda_c\widetilde{g}_{ab}^*f_{bc}^*+(1-\Lambda_c)\,\widetilde{g}_{ac}^*\right)\cdot\left(\alpha_c\,(1-\Lambda_c)+\alpha_y\Lambda_yf_{yc}\right)\right.$$
$$\left.+\left(\alpha_b^*\Lambda_bf_{bc}^*+\alpha_c^*\,(1-\Lambda_c)\right)\cdot\alpha_x\left((1-\Lambda_c)\,\widetilde{g}_{xc}+\Lambda_c\widetilde{g}_{xy}f_{yc}\right)\right]$$
$$+\epsilon^2\alpha_a^*\alpha_x\left[\left(\widetilde{g}_{ab}^*f_{bc}^*+(1-\Lambda_c)\,\Gamma_{ac}^*\right)\cdot\Lambda_c\cdot\left((1-\Lambda_c)\,\Gamma_{xc}+\widetilde{g}_{xy}f_{yc}\right)\right]$$

where we have explicitly separated out the different orders in ϵ to show the unperturbed, interference, and perturbation effects. If $P_3 = P_1$ this is the entire power; otherwise, we must add the P_{3+} and P_{1+} terms as mentioned.

Interestingly, only one factor of $1/\Lambda$ remains, and at $\mathcal{O}(\epsilon^2)$.

As the formula is complicated, we show the $P_3 = P_1$ field and power for the APLC

$$Field = \alpha_a(1-\Lambda_a)P_1\left|a\right\rangle + \epsilon\cdot\alpha_a\Gamma_{ab}(1-\Lambda_b)P_1\left|b\right\rangle$$

$$Power = \Lambda_c(1-\Lambda_c)^2\left(|\alpha_c|^2 + \epsilon\cdot[\alpha_a^*\alpha_c\Gamma_{ac}^* + \alpha_c^*\alpha_a\Gamma_{ac}] + \epsilon^2\cdot\alpha_a^*\alpha_b\Gamma_{ac}^*\Gamma_{bc}\right)$$

still summing over a, b, c. As noted, the Γ matrices from the perturbation simply shift the α coefficients.

We now can consider the residual either as the ratio of the Lyot plane power, (4.42), to the perturbed power through the pupil plane (4.40), or study its ratio with the unperturbed pupil plane power $\sum_a |\alpha_a|^2\Lambda_a$ (4.11). If the perturbation is small enough, then we can expand in a Taylor series around it. The former ratio then symbolically becomes

$$\frac{A+\epsilon B+\epsilon^2 C}{D+\epsilon E+\epsilon^2 F} \approx \frac{A}{D} + \epsilon\left(\frac{B}{D} - \frac{AE}{D^2}\right) + \epsilon^2\left(\frac{C}{D} - \frac{BE+AF}{D^2} + \frac{AE^2}{D^2}\right) + \dots$$

to second order in epsilon, which gives us the additional power corrections that dividing by the unperturbed pupil-plane power (D in the shorthand) would miss.

The formula for the throughput, likewise, can compare the pupil-plane photon rate to the perturbed photon rate through the unapodized pupil or the unperturbed photon rate through the unapodized pupil. If the former, then the change for the perturbation will be encoded both in the additional terms from (4.40) relative to (4.11) and from the shift in the maximum magnitude

in the pupil that the perturbation causes; the latter case will have additional terms from the denominator.

Let's examine photon rate through the unapodized pupil; this is just

$$\frac{1}{(\mathcal{F}')^2}\frac{\sum|\alpha_a|^2\Lambda_a + A\epsilon + B\epsilon^2}{4\pi\sum\Lambda_a}$$

using quick shorthand for the additional power terms. $\mathcal{F}'$ is the new maximum absolute value of the perturbed field $(1+\epsilon g)\,|\alpha\rangle$. We can show that if the perturbation only moves the maximum by a small distance and small amount – i.e. the non-analytic behavior caused by the sharp boundaries of the secondary does not occur – then

$$\mathcal{F}'^2 = \mathcal{F}^2\left[1+\epsilon(g+g^*)+\epsilon^2\left(|g|^2 - \frac{1}{2}\sum(\partial_j g)^2\frac{\mathcal{F}}{\partial_j^2\Phi} - \frac{1}{2}\,c.c.\right)\right]$$

with all terms evaluated at the point of the old maximum, $\Phi = \sum_a \alpha_a\phi_a$, and the local coordinates j are chosen so that the mixed derivatives of Φ vanish. The throughput therefore takes on the same generic form as the residual energy (ratio of quadratic polynomials in ϵ), and may also be expanded in ϵ. We have not found the explicit forms to show useful information, and so neglect to write them here.

Recall that in (4.17) we demonstrated the ability to calculate the instrument-plane intensity at $\rho = 0$. There, we used the approximation that the off-axis peak was the maskless on-axis peak. We can now use the same method to calculate the $\rho = 0$ instrument plane intensity ratio for the perturbed and unperturbed light. So long as the perturbation does not move the light's maximum off-axis, this will be the Strehl ratio. We will do so under the

$P_3 = P_1$ approximation; the full formula can be found by applying $\langle \rho = 0|$ to (4.41) to capture the P_{3+} and P_{1+} effects.

We re-use the fact that $\langle \rho = 0| P_1 |c\rangle = \Lambda_c \sum_{t\ even} V_{c,t0} \sqrt{\frac{t+1}{\pi}}$, which gives us

$$S = \left| 1 + \epsilon \cdot \frac{\sum \alpha_a \widetilde{g}_{ab} [\delta_{bc} + \Lambda_c (f_{bc} - \delta_{bc})] V_{c,t0} \sqrt{\frac{t+1}{\pi}}}{\sum \alpha_a \Lambda_a [\delta_{ac} + \Lambda_c (f_{ac} - \delta_{ac})] V_{c,t0} \sqrt{\frac{t+1}{\pi}}} \right|^2 \tag{4.43}$$

though this neglects the fact that the perturbed wave is slightly more powerful than the unperturbed one in the $1 + \epsilon g$ form that we have been using. We can correct for this if desired by dividing the numerator by the square root of (4.40) and the denominator by $\sqrt{\sum |\alpha_a|^2 \Lambda_a}$. Doing so and expanding to first order in ϵ overall, we have

$$S = 1 + 2\epsilon \,\mathrm{Re} \left\{ \frac{\sum \alpha_a \widetilde{g}_{ab} [\delta_{bc} + \Lambda_c (f_{bc} - \delta_{bc})] V_{c,t0} \sqrt{\frac{t+1}{\pi}}}{\sum \alpha_a \Lambda_a \delta_{ab} [\delta_{bc} + \Lambda_c (f_{bc} - \delta_{bc})] V_{c,t0} \sqrt{\frac{t+1}{\pi}}} - \frac{\frac{1}{2} \sum \alpha_a^* (\widetilde{g}_{ac} + \widetilde{g}_{ac}^*) \alpha_c}{\sum \alpha_a^* \Lambda_a \delta_{ac} \alpha_c} \right\} \tag{4.44}$$

As a check, a constant $g = 1$ implies that $\widetilde{g} = K$, the kernel. Choosing a single $\alpha_a = 1$ and setting $f = 0$ (APLC), the quantity in braces simplifies to $\frac{\Lambda_a}{\Lambda_a} - \frac{\Lambda_a}{\Lambda_a} = 0$ so that merely changing the intensity of the light does not change a power-normalized intensity ratio. If we wish to neglect the normalization by the power, then the second term in the brackets is removed, and the ratio has changed to $1 + 2\epsilon$ to lowest order, as would be expected.

4.4 Summary

The use of Slepian modes in propagating light through the coronagraph has given us a number of benefits in monochromatic propagation. Even with a linear sum of these eigenfunctions for the apodization, which includes the clear pupil, we find relatively simple expressions for the photon throughput and Lyot plane residual energy. For the π phase mask and APLC, these expressions show that when the Lyot stop is equal to the pupil the residual energy is a weighted sum of that for the modes involved. The throughput is nearly so; improvement only occurs if the combined apodization is flatter than the main modes involved.

A change of wavelength $\lambda_{old}/\lambda_{new} = \eta$ causes a rescaling $r \to \eta r$. The recursion relations of the Bessel functions mean that this action can also be represented as a matrix operator, referred to as the dilation operator. This means that broadband analysis of the Slepian modes and eigenvalues can be carried out via matrix transforms after the kernel is calculated once, with a somewhat larger number of basis functions than would be strictly necessary otherwise. The dilation operator is easily capable of handling 25% bandwidths.

For the calculation of the different-wavelength apodizations, we recommend using $\eta < 1$, though $\eta \lesssim 1.1$ is not impractical. Therefore, the direct kernel calculation should be done at the shortest wavelength. However, the dilation operator also allows us to follow the propagation of a plane wave through apodizations designed around a different wavelength. In this case, we wind up requiring dilations in $1/\eta$. This would put a lower limit on $0.9 \lesssim \eta$, which still allows $20 - 25\%$ bandwidths. It may be best to perform

the kernel calculations twice, bracketing the bandwidth. Doing so will give us the greatest ability to track instrument response, as well as provide a measure of error checking.

We have a wholly accurate, simple expression to calculate the derivative of the eigenvalues and eigenvectors with respect to $\mathcal{N}$. We can therefore quickly determine a part of the chromatic dispersion of the system. This derivative expression also allows us to prove that this derivative must be positive or zero, so that all eigenvalues must increase as mask size does. Coronagraphs with wavelength-dependent mask or apodization responses will require more careful analysis than we have performed.

The recursion relations also allow us to find simple matrices for the action of polynomials $r^n e^{i\ell\theta}$ on pupil apodizations. Consequentially, we can follow slowly-varying perturbations away from a plane wave without introducing an excessive number of additional basis functions. The practical limit is likely $\sim r^5$, though we have explored r^8. Small-scale variations will not be handled well by this formalism.

Slightly off-axis illumination can be handled by the polynomials r^n. The polynomials cannot handle strongly off-axis illumination. A peculiarity of the Slepian duality gives us at least an approximation for this field. If necessary, we can create the instrument-plane field by shifting the on-axis, image-plane, mask-limited field. This will contain a fraction Λ of the total energy for the off-axis source, though it will neglect any of the diffraction pattern that occurred off-mask. Overlap with the mask can be written in the eigenvectors and therefore adjusted for in the Lyot and instrument planes.

Future possibilities

The four-quadrant phase mask and vortex coronagraphs rely on redirection of light to the outer edges of the Lyot stop through phase manipulation, rather than cancellations within the Lyot stop. In exploring our basis functions previously, we found that the mathematical mechanism by which these occur: "basis modes" with $m > t$. These "modes" leave a perfectly dark core when they are not limited by pupil edges.

With the class of such functions determined, it is interesting to speculate on other uses for them. They naturally will be limited by chromatic effects, as they can never be wholly positive. We wonder if the resulting problems would be outweighed by the benefits of the dark core remaining after limitation to the pupil, and its behavior.

For the bandpass masks of (Kuchner and Traub, 2002), we have found that expanding the $1 - \sin$, $1 - J_0$, and $\sin^2$ masks to $R_4^0(\rho)$ is usually sufficient to reproduce them. In so doing, their coefficients are all fairly similar. We speculate it may be possible to optimize this style of mask by tweaking these coefficients rather than searching for different similar functions.

We also call attention to our general formula for energy within a given radius $\rho_0 = (2/\mathcal{N})\zeta_0$ in the instrument plane, and the averaged intensity on the edge thereof. While we have not developed it further in for arbitrary pupils, we will see in chapter 5 that the circular pupil case shows remarkable simplification. It may be that our neglect of the general case is mistaken, and that treating results as perturbations from the circular-pupil case is within practical use.

Proof: Shift of maximum value

Let $f(x)$ and $g(x)$ be complex, with the maximum $|f| \equiv \mathcal{F}$ occurring at x_1 and the maximum of $f + \epsilon g$ occurring at x_2. We will adapt an overall global phase so that $f(x_1) = \mathcal{F}$.

The value $f(x_2) = \mathcal{F} + \frac{1}{2}\sum(x_2 - x - 1)_j(x_2 - x_1)_k\partial_j\partial_k f$ by expanding around x_1. Likewise, $g(x_2) = g + (x_2 - x_1)_j\partial_j g$.

We had that $\nabla_2(f + \epsilon g) = 0$, so this must also be equal to $\sum_j(x_2 - x_1)_j\partial_j\partial_k f + \epsilon\partial_k g$. Assume coordinate axes so that the mixed derivative is zero at this point. Requiring the two terms to cancel, as both ϵ and $(x_2 - x_1)$ are small, yields $(x_2 - x_1)_j = -\epsilon(\partial_j g)/(\partial_j^2 f)$ to lowest order approximation.

This means that at x_2, $f \approx \mathcal{F} + \frac{1}{2}\epsilon^2\sum\frac{(\partial_j g)^2}{\partial_j^2 f}$. Similarly, $g(x_2) \approx g - \epsilon\sum(\partial_j g)^2/(\partial_j^2 f)$ The squared magnitude at x_2 is therefore

$$\mathcal{F}^2 + \epsilon\mathcal{F}(g + g^*) + \epsilon^2\left(|g|^2 - \frac{1}{2}\mathcal{F}\sum\frac{(\partial_j g)^2}{\partial_j^2 f}\right)$$

Returning to the field perturbation, there is a slight change of nomenclature; we are dealing with $(1 + \epsilon g)f$. Substitute $g \to gf$. Since we are at x_1, $\partial_j(gf) = \mathcal{F}\partial_j g$. This gives the final form

$$\mathcal{F}'^2 = \mathcal{F}^2\left[1 + \epsilon(g + g^*) + \epsilon^2\left(|g|^2 - \frac{1}{2}\sum(\partial_j g)^2\frac{\mathcal{F}}{\partial_j^2 f} - \frac{1}{2}\,c.c.\right)\right]$$

where all functions and derivatives are evaluated at x_1. The term in ϵ and the first term in ϵ^2 are just the influence adding g had on the maximum at x_1. The second term in ϵ^2 represents the influence of the change of the location, balancing the gradient in g with the curvature in f.

Note on numerical transforms

We make here notes about two brief effects within our framework relating to these numeric transforms. The first is on frequency spacing in the array. The second is on improved approximations.

Let us pixellate the Lyot stop and use a periodogram (FFT) to rapidly calculate the instrument-plane fields in an array of the same size. For arrays of size $W \times W$, the phase of the Fourier integral $ir\rho$ corresponds (up to overall shifts and zeroing) to the phase of the discrete Fourier sum $i\pi jk/W$. j and k are the indices for r and ρ. If this $W \times W$ array aligns exactly with the edge of the pupil, then $r_j \sim \mathcal{N}\pi/2 \times j/(W/2)$. Putting this into the phase, we find that $k \sim \mathcal{N}\rho_k$. We rewrite this in terms of ζ, our measure for angular widths in units of λ/D_P, as $\rho = \zeta 2/\mathcal{N}$.

The result is that for such an FFT approximation, the array spacing in the instrument plane will be $\frac{1}{2}\lambda/D_P$. This may be sufficient for some needs, allowing us to bypass lengthy numerical integrations.

Our second brief note is on improving intensity approximations. Let us look at the power that will illuminate a CCD. As this is a rough approximation to illustrate the concept, we will say that the CCD is a circle of radius p, as measured in $\rho-$coordinates ($\mathcal{N}\pi p/2$ in λ/D_P). The power is therefore

$$\int_{CCD} \mathrm{d}^2 q \, |\Phi_D|^2 = \int_{CCD} \mathrm{d}^2\rho \int_{P3} \mathrm{d}^2 r \, \mathrm{d}^2 r' \sum_{jk} C_{jk} [b_j(r,\theta)]^* \, b_k(r',\theta') e^{i(\boldsymbol{\rho}_0+\mathbf{q})\cdot(\mathbf{r}'-\mathbf{r})}$$

C_{jk} are the resulting coefficients from expanding the Slepian functions in the basis and allowing for mask effects. $\boldsymbol{\rho}_0$ is the center of the pixel.

We can expand the exponential and perform the q integral. Using the circular approximation for the CCD means that the first order term of the expansion, $\int \mathrm{d}^2 q\, \mathbf{q} \cdot (\mathbf{r}' - \mathbf{r})$ is zero, and the leading approximation is therefore $\pi p^2 \left(1 - p^2 |\mathbf{r}' - \mathbf{r}|^2/8 + \ldots\right)$.

We have already proven § 3.3.3 that r^2 and $re^{i\theta}$ can be expressed as linear operators on the basis functions. This means that we can take these linear combinations outside of the integrals over r and r', and combine them with the C_{jk} coefficients already present. The approximate detector power is then

$$\pi p^2 \left(|\Phi_D(\boldsymbol{\rho}_0)|^2 - \frac{p^2}{8} \sum_{j,k,x,y} C_{jk} \left[(r^2)_{jx}\delta_{ky} + (r^2)_{ky}\delta_{jx} - 2\,\mathrm{Re}\left\{ (re^{i\theta})_{jx} (re^{i\theta})_{ky} \right\} \right] \right.$$
$$\left. \times \int_{P3} \mathrm{d}^2 r\, (b_x(r))^* e^{-i\boldsymbol{\rho}_0 \cdot \mathbf{r}} \int_{P3} \mathrm{d}^2 r'\, b_y(r) e^{i\boldsymbol{\rho}_0 \cdot \mathbf{r}'} \right)$$

Dividing by the detector area gives us an improved approximation for the average intensity at that location. We have not pursued this possibility.

References

Kuchner, Marc J. and Wesley A. Traub (2002). "A Coronagraph with a Band-limited Mask for Finding Terrestrial Planets". In: *ApJ* 570.2.

Soummer, R., L. Pueyo, A. Ferrari, C. Aime, and A. Sivaramakrishnan (2009). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures, II. Theoretical Properties and Application to Extremely Large Telescopes". In: *The Astrophysical Journal* 695.1, pp. 695–706.

NIST Digital Library of Mathematical Functions. "http://dlmf.nist.gov/, Release 1.0.14 of 2016-12-21". URL: "http://dlmf.nist.gov/".

Krist, John E., Ruslan Belikov, Laurent Pueyo, Dimitri P. Mawet, Dwight Moody, John T. Trauger, and Stuart B. Shaklan (2011). *Assessing the performance limits of internal coronagraphs through end-to-end modeling: a NASA TDEM study*. DOI: 10.1117/12.892772. URL: https://doi.org/10.1117/12.892772.

Laurent, Kathryn St., Kevin Fogarty, Neil T. Zimmerman, Mamadou N'Diaye, Christopher C. Stark, Johan Mazoyer, Anand Sivaramakrishnan, Laurent Pueyo, Stuart Shaklan, Robert Vanderbei, and Rèmi Soummer (2018). *Apodized pupil Lyot coronagraphs designs for future segmented space telescopes*. DOI: 10.1117/12.2313902. URL: https://doi.org/10.1117/12.2313902.

Ruane, Garreth, A Riggs, Johan Mazoyer, Emiel Por, Mamadou N'Diaye, Elsa Huby, P Baudoz, R Galicher, Ewan Douglas, Justin Knight, Brunella Carlomagno, K Fogarty, L Pueyo, N Zimmerman, Olivier Absil, Mathilde Beaulieu, E Cady, Alexis Carlotti, David Doelman, and M Ygouf (2018). *Review of high-contrast imaging systems for current and future ground- and space-based telescopes I. Coronagraph design methods and optical performance metrics*.

Leboulleux, Lucie, Laurent Pueyo, Thierry Fusco Jean-FranÃğois Sauvage, Johan Mazoyer, Anand Sivaramakrishnan, Mamadou N'Diaye, and Rèmi Soummer (2018). *Sensitivity analysis for high-contrast imaging with segmented*

space telescopes. DOI: 10.1117/12.2313904. URL: https://doi.org/10.1117/12.2313904.

Born, Max and Emil Wolf (1999). *Principles of Optics. Seventh (Expanded) Edition*. Cambridge, UK: Cambridge University Press.

Chapter 5

Circular Pupil Slepians and Non-Circular Demonstration

We now apply the results found in chapter 3 to two geometries at a variety of wavelengths. Our focus is primarily on proof of concept, though we are still able to draw general conclusions.

§ 5.1 will focus on creation of the Slepian modes for circular geometries. This will include central obstructions of different sizes. As the generation of kernel elements, equation (3.15), is integrated over area, the differences between more complex pupils and this geometry go roughly as the relative area of the difference.

To begin, we will discuss several simplifications which occur in this highly symmetric case. Notably, we have an explicit formula for the elements of the kernel, which decomposes into different blocks of fixed angular mode. This allows us to strictly order eigenvalues belonging to the even mode, and separately to the odd mode. We also show that there is a simple formula for the angularly averaged instrument plane intensity, though it still relies on

numerical integrals.

Following the analytical work, we will first compare our results to those found previously for such an APLC (Soummer et al., 2009), and show that our results are wholly in agreement. Once the validity is established, we turn to the behavior of the other modes under parameter changes, though we leave aside propagation through the coronagraph. We also demonstrate that we can explain the "bell-bagel" transition in our framework.

Following the discussion of the Slepian modes of circular pupils, we generate a hypothetical APLC for the James Webb Space Telescope § 5.2. Since this is not true design work, but proof of concept, larger relative errors were tolerated. We will first review the relevant parameters of the JWST, and show that these allow an approximation of the instrument response using discrete Fourier transforms to within satisfaction for our purpose here.

The Slepian modes and their resulting PSFs are then explored. After they are established, we demonstrate our ability to find positive-only apodizations using sums of the modes. This will allow optimization of instrument-plane contrast levels by altering the α coefficients, a substantially reduced space to probe compared to pointwise designs. We then exhibit the behavior of the modes when generated for a wide band of wavelengths using the dilation method, and show that our labelling of modes by eigenvalue rank a obscures a continuity.

Our conclusions are gathered in § 5.3

5.1 Circular APLC: Prior and New Results

5.1.1 Analytical Simplifications for Circular Pupils

The symmetry of the pupil allows us several well-known simplifications; most notably, the complete separation of the angular and radial modes for inner products on the pupil. The kernel immediately factorizes, with each element $K_{tm,sn} = K_{ts}\delta_{mn}$. Each Slepian mode therefore can be categorized by an angular and radial mode number, and we can discuss the behavior of decreasing-eigenvalues within a given angular mode.

This factorization means that only the $m = 0$ block can have an eigenfunction which is wholly positive on the pupil, as all others have factors of $e^{im\theta}$. This is essentially a one-dimensional problem. Since the eigenfunctions are still orthogonal, and it is impossible for the integral of the product of two positive functions to be zero, at most one of the eigenfunctions can be wholly positive.

That this wholly-positive function corresponds to the highest eigenvalue is a theorem from the study of Sturm-Liouville theory. While we have not proven to ourselves that our integral equations are the equivalent of such a problem, we remain fairly confident that this is the case. We do not believe that this positivity condition extends to the general two-dimensional case. Section 5.2 will show that the JWST pupil does not have an all-positive mode.

The symmetrical simplification also causes each $m-$block of the kernel to be a truncation of the $m-2$ block before it, as the same radial integrals are performed but now exclude the old $t = m-2$ entries. Cauchy's interlacing

theorem then applies, which means that the eigenvalues of the higher $m-$block alternate with those of the lower $m-2$ block:

$$\Lambda_{1,m-2} \geq \Lambda_{1,m} \geq \Lambda_{2,m-2} \geq \Lambda_{2,m-2} \geq \dots$$

This interleaving continues, since the $m-4$ block is a subset of the $m-2$ block and so on. As a consequence,

$$\Lambda_{1,m=0} \geq \Lambda_{1,m=2} \geq \Lambda_{2,m=0} \geq \Lambda_{1,m=4} \geq \Lambda_{2,m=2} \geq \Lambda_{3,m=0} \geq \dots > 0$$

and similarly for the odd-m eigenvalues. However, we cannot say how the odd and even m eigenvalues compare to each other.

The radial integrals of the kernel elements,

$$2\sqrt{(t+1)(s+1)} \int \mathrm{d}r \, \frac{J_{t+1}(r) \, J_{s+1}(r)}{r}$$

are exact analytical functions.

$$K_{t,s\neq t} = 2\sqrt{(t+1)(s+1)} \times \left[\frac{r[J_s(r) J_{t+1}(r) - J_t(r) J_{s+1}(r)] - (s-t)[J_{s+1}(r) J_{t+1}(r)]}{(s+1)^2 - (t+1)^2} \right]$$

$$K_{t,t} = \left[1 - J_0(r)^2 - J_{t+1}(r)^2 - 2\sum_{j=1}^{t} J_j(r)^2 \right] \tag{5.1}$$

We take the difference of these functions evaluated at $r = \mathcal{N}\pi/2$ and $r = R_S\mathcal{N}\pi/2$ to calculate the definite integral for the kernel. We have used (*NIST Digital Library of Mathematical Functions*) 10.22.6 for the $t \neq s$ case; the $t = s$ case uses the standard recursion formula $J_{n+1}(r) + J_{n-1}(r) = \frac{2n}{r} J_n(r)$ with

10.22.30 from the same. This explicit formulation, combined with the angular-mode factorization, allows us to rapidly generate the kernel for even a very large number of basis functions.

The circular pupil also aids us at the Lyot plane. All integrals over regions P_{1+} and P_{3+} are over annular regions. The same factorization between angular and radial modes continues to apply, and integrals of products of jinc functions still follow (5.1), though now evaluated between different limits. Moreover, so long as we assume standard undersizing — $R_{L,in} \geq R_S$ and $R_{L,out} \leq R_P$ — then all expressions involving P_{3+} in chapter 3 drop out.

We are not aware of any general formula for the transform from the Lyot plane to the instrument plane, whose integral reduces to

$$\int \mathrm{d}r\, r \mathcal{J}_{t+1}(r)\, J_m(\rho r)$$

up to overall constant factors and $e^{im\varphi}$. The growing error in results of dilation by very small η in 4.2 mean that we should not approach this using $\mathcal{J}_{t+1}(r) = [D_{\rho^{-1}}]_{ts}\mathcal{J}_{s+1}(\rho r)$, even if it is mathematically correct.

However, the angularly-averaged intensity in the instrument plane for a single mode, (4.10), now enjoys a considerable simplification.

This is sufficiently interesting to extend it to the general $|\alpha\rangle$ and f. The extension can be carried further, if we recall that the action of perturbations

was to shift $\alpha \to \alpha(1+\epsilon\Gamma)$, but we will not write it here explicitly.

$$\langle I_D(\rho_0)\rangle = \frac{1}{2\pi\rho_0}\frac{\mathrm{d}}{\mathrm{d}\rho_0}\int_0^{\rho_0} \mathrm{d}^2\rho\,|\Phi_D|^2$$

$$= \frac{1}{2\pi\rho_0}\frac{\mathrm{d}}{\mathrm{d}\rho_0}\sum_{abcd}\alpha_a^*\alpha_c[(1-\Lambda_a)\delta_{ab}+\Lambda_a f_{ab}^*][(1-\Lambda_c)\delta_{cd}+\Lambda_c f_{cd}]$$

$$\times \sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'}\int \mathrm{d}^2\rho \int_{P3}\mathrm{d}^2 r \int_{P3}\mathrm{d}^2 r'\, b_{tm}^*(\mathbf{r}) b_{t'm'}(\mathbf{r}') e^{i\boldsymbol{\rho}\cdot(\mathbf{r}-\mathbf{r}')}$$

$$= \frac{1}{2\pi\rho_0}\frac{\mathrm{d}}{\mathrm{d}\rho_0}\sum_{abcd}\alpha_a^*\alpha_c[\sim][\sim]\sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'}\int_{P3}\mathrm{d}^2 r \int_{P3}\mathrm{d}^2 r'\, b_{tm}^*(\mathbf{r}) b_{t'm'}(\mathbf{r}')$$

$$\times 4\pi\rho_0^2\sum_{k=0}^{\infty}\sum_{\substack{j=-k\\ \mathrm{j{+}{+}}}}^{k}(k+1)\mathcal{J}_{k+1}(\rho_0 r)\,\mathcal{J}_{k+1}(\rho_0 r')\,e^{ij(\theta-\theta')}$$

$$= \frac{1}{\rho_0}\sum_{abcd}\alpha_a^*\alpha_c[\sim][\sim]\sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'}\sum_{k=0}^{\infty}\sum_{\substack{j=-k\\ \mathrm{j{+}{+}}}}^{k} 4(k+1)\rho_0$$

$$\times\Big[\int_{P3}\mathrm{d}^2 r\,\left(J_k(\rho_0 r)-J_{k+2}(\rho_0 r)\right) b_{tm}^*(\mathbf{r})e^{ij\theta}\int_{P3}\mathrm{d}^2 r'\,\mathcal{J}_{k+1}(\rho_0 r')\,b_{t'm'}(\mathbf{r}')e^{-ij\theta'}$$

$$+\int_{P3}\mathrm{d}^2 r\,\mathcal{J}_{k+1}(\rho_0 r)\,b_{tm}^*(\mathbf{r})e^{ij\theta}\int_{P3}\mathrm{d}^2 r'\,\left(J_k(\rho_0 r')-J_k(\rho_0 r')\right)b_{t'm'}(\mathbf{r}')e^{-ij\theta'}\Big]$$

$$= 8\pi \sum_{abcd} \alpha_a^* \alpha_c [\sim][\sim] \sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'} \sqrt{(t+1)(t'+1)} \sum_{k=0}^{\infty} \sum_{\substack{j=-k \\ \text{j++}}}^{k} \delta_{jm} \delta_{jm'}$$

$$\left[\int_{P3} \mathrm{d}r \, \left(J_k\left(\rho_0 r\right) - J_{k+2}\left(\rho_0 r\right) \right) J_{t+1}\left(r\right) \int_{P3} \mathrm{d}r' \, \left(J_k\left(\rho_0 r'\right) + J_{k+2}\left(\rho_0 r'\right) \right) J_{t'+1}\left(r'\right) \right.$$

$$\left. + \text{unprimed} \leftrightarrow \text{primed} \right]$$

$$= 16\pi \sum_{abcd} \alpha_a^* \alpha_c [\sim][\sim] \sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'} \sqrt{(t+1)(t'+1)} \sum_{k=0}^{\infty} \sum_{\substack{j=-k \\ \text{j++}}}^{k} \delta_{jm} \delta_{jm'}$$

$$\times \int_{P3} \mathrm{d}r \int_{P3} \mathrm{d}r' \, \left[J_k\left(\rho_0 r\right) J_{t+1}\left(r\right) J_k\left(\rho_0 r'\right) J_{t'+1}\left(r'\right) \right.$$

$$\left. - J_{k+2}\left(\rho_0 r\right) J_{t+1}\left(r\right) J_{k+2}\left(\rho_0 r'\right) J_{t'+1}\left(r'\right) \right]$$

Because of the interaction of the delta functions and the $k - (k+2)$ terms, all k except $k = m = m'$ cancel or are zero. Therefore,

$$\begin{aligned} \langle I_D(\rho_0) \rangle = 16\pi & \sum_{abcd} \alpha_a^* \alpha_c [(1-\Lambda_a)\delta_{ab} + \Lambda_a f_{ab}^*][(1-\Lambda_c)\delta_{cd} + \Lambda_c f_{cd}] \\ & \times \sum_{tm,t'm'} V_{b,tm}^* V_{d,t'm'} \sqrt{(t+1)(t'+1)} \delta_{mm'} \\ & \times \int_{P3} \mathrm{d}r \, J_m\left(\rho_0 r\right) J_{t+1}\left(r\right) \int_{P3} \mathrm{d}r' \, J_{m'}\left(\rho_0 r'\right) J_{t'+1}\left(r'\right) \end{aligned} \quad (5.2)$$

The integrals are from the inner Lyot stop to the outer Lyot stop, and must be done numerically as no satisfactory analytic solution exists.

The power within ρ_0 can be found by multiplying by $2\pi\rho_0$ before integrating $\int \mathrm{d}\rho_0$, with use of the identities (*NIST Digital Library of Mathematical*

Functions) 10.22.4 and 5:

$$\int \mathrm{d}z\, z J_\mu(az)\, J_\mu(bz) = \frac{z\left(aJ_{\mu+1}(az)\, J_\mu(bz) - J_\mu(az)\, J_{\mu+1}(bz)\right)}{a^2 - b^2} \qquad a \neq b$$

$$= \tfrac{1}{2}z^2\left(\left[J_\mu(az)\right]^2 - J_{\mu-1}(az)\, J_{\mu+1}(az)\right) \qquad a = b$$

5.1.2 Comparison to Prior Results

In comparing our results, the basis functions used were all t up to the lesser of $2.5 \times \mathrm{Tr}\,(K)$ or 20. As a consistency check, the sum of the calculated eigenvalues was compared to the predicted value for $\mathrm{Tr}\,(K)$; the ratio was found to differ from unity by $10^{-2} - 10^{-6}$ over the range calculated, with the most significant deviations occurring in the region $\mathcal{N} \in (2.5, 3.5)$ and $R_S > 0.35$.

Figure 5.1 is a comparison of the normalized Lyot-plane residual energy, $(1 - \Lambda_a)^2$, from (Soummer et al., 2009) to our results; they quite clearly match. Likewise, we compare the throughputs in figure 5.2. While there is a slight mismatch in the $\mathcal{N} \approx 7$, $R_S \approx 0.4$ region, we are not concerned given the exceptional match otherwise shown.

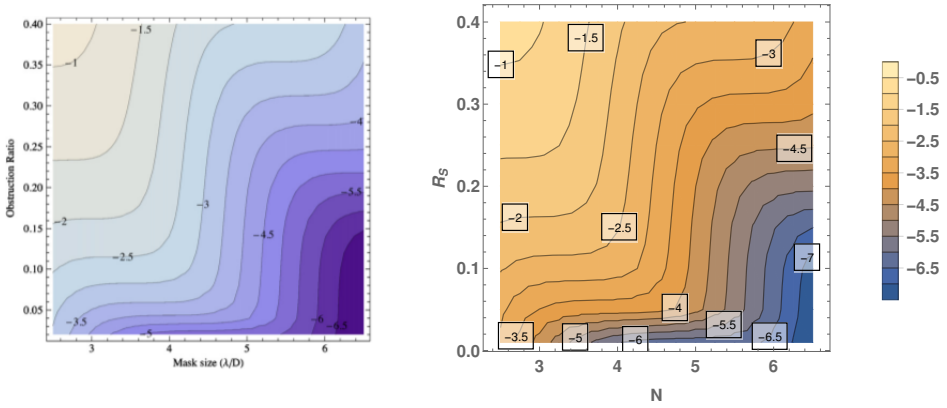


Figure 5.1: $\log_{10}$ residual energy from (Soummer et al., 2009) (left), and our method (right).

If we look ahead slightly to figure 5.11, we see that there are regions of parameter space where the largest eigenvalue is that with $|m| = 1$, instead of $m = 0$. This eigenfunction is necessarily not strictly positive, since it is multiplied by a sine or cosine.

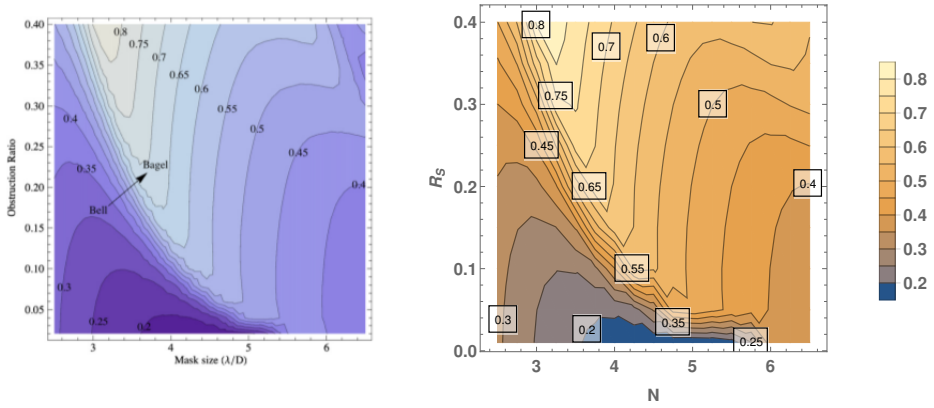


Figure 5.2: $\log_{10}$ residual energy from (Soummer et al., 2009) (left), and our method (right).

This poses a challenge; the iterative procedure described in (Soummer et al., 2009) was predicated on the belief that Λ_1 corresponded to the positive mode. This is clearly not the case, despite the excellent correspondence to the $m = 0$ results. This is concerning, as the development of non-symmetric pupils has proceeded without problem.

The resolution comes from the initial choice in that algorithm, which was the constant pupil function. In section 4.3.1 we used the mathematical identity

$$1 = \sum_{\substack{t=0 \\ \text{t even}}}^{\infty} 2(t+1)\mathcal{J}_{t+1}(r)$$

This is entirely within the subspace spanned by even t, and therefore is incapable of including the odd t needed to develop the $|m| = 1$ mode. Solutions will therefore converge to the highest even-only result, which is the largest-Λ $m = 0$ solution thanks to the Cauchy interweaving.

For non-circular pupils, the block structure of the kernel is broken and we do not have the separation of the even- and odd-t modes. The choice of a

constant pupil will regain its overlap with the largest Λ, and the algorithm will converge, though it will be slow in those regions where the base mode is primarily composed of odd-m basis functions.

5.1.3 Angular modes

We now turn to the results for the full range of modes, which we study over a larger range of parameters. A representative set of these Slepian modes are shown in figure 5.3 for a mask of $\mathcal{N} = 3.5$ and an obstruction of $R_S = 20\%$. These parameters fall in one of several regions where the highest $m = 1$ mode is of higher eigenvalue than that of the $m = 0$ mode.

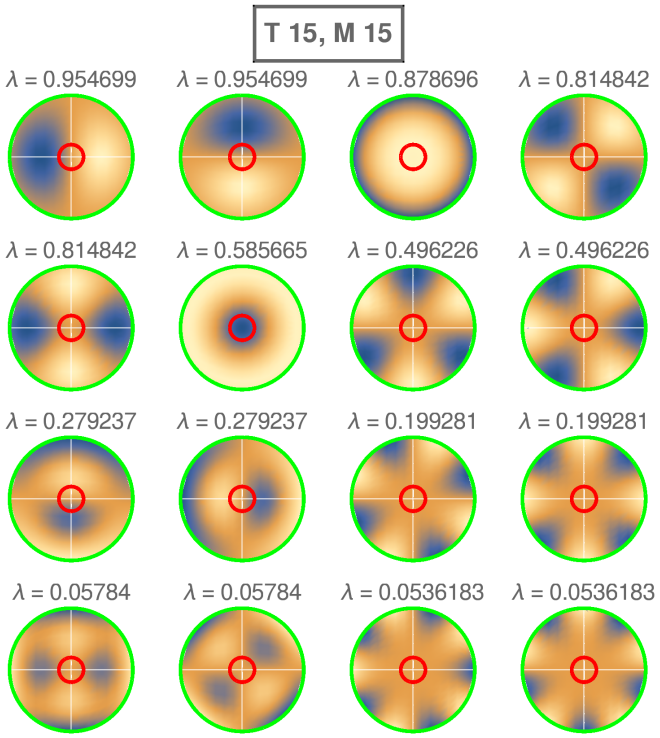


Figure 5.3: Density plots of the top twelve ϕ_a for $\mathcal{N} = 3.5$ and $R_S = 0.2$. White-orange is the maximum value in each plot; blue is the minimum value, which is negative for all but the $m = 0$, $k = 0$ case.

The expanded throughput and residual for the highest $m = 0$ mode is shown in 5.4; the second $m = 0$ mode is shown in 5.5. Results for all $|m| > 0$ are doubly degenerate, so we show the results for the two highest unique results. $|m| = 1$ are in figures 5.6 and 5.7; $|m| = 2$ are in figures 5.8 and 5.9.

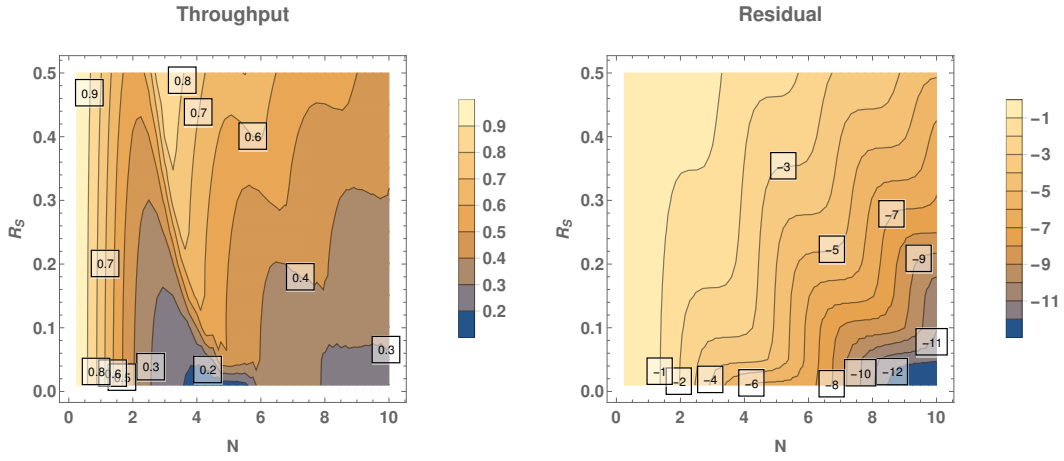


Figure 5.4: Contour plots of throughput and (log) residual energy for the first $m = 0$ mode.

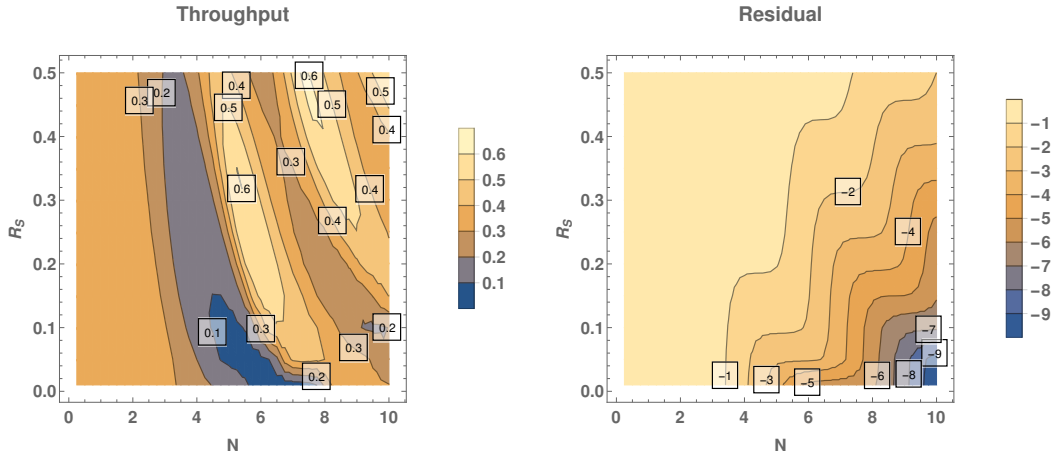


Figure 5.5: Contour plots of throughput and residual energy for the second $m = 0$ mode.

In figure 5.10, we overlay the contour lines from the residuals, which

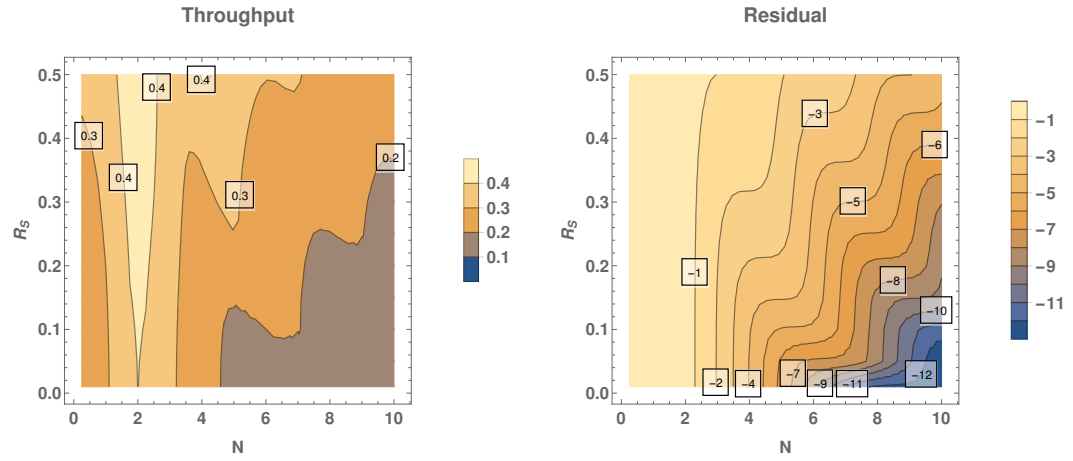


Figure 5.6: Contour plots of throughput and residual energy for the first (two degenerate) $|m| = 1$ modes.

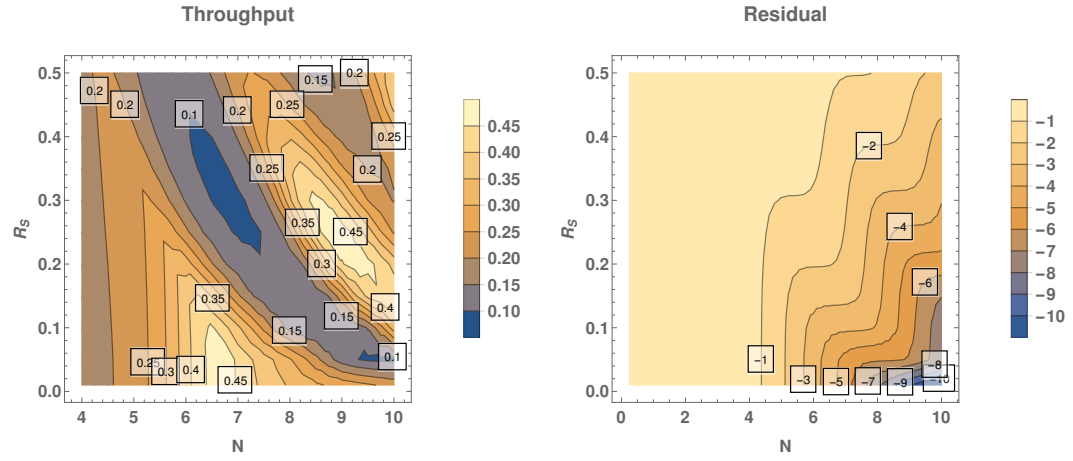


Figure 5.7: Contour plots of throughput and residual energy for the second (two degenerate) $|m| = 1$ modes.

shows the alternating locations of the ledges and cliffs of the contour plot. This intersection means that a cross-section at fixed R_S, figure 5.11, shows several interesting behaviors. At the center of a ledge for a fixed m up to 4, both the $m-1$ and $m+1$ modes have their largest $\frac{\partial e}{\partial N}$, as well as obeying $e_m = e_{m-1} = e_{m+1}$. Each $\log_{10} e_m$ also seems to follow a similar function. The $k = 2$ modes obey a similar relation.

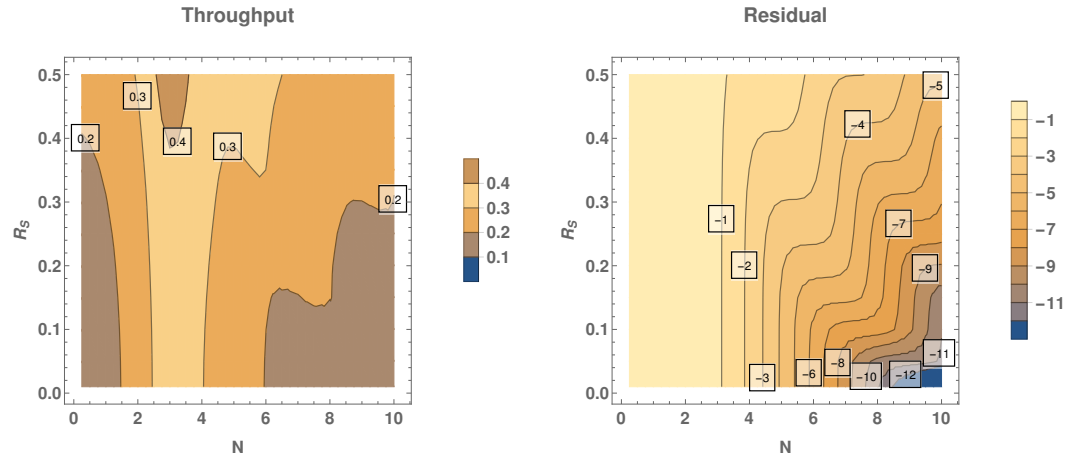


Figure 5.8: Contour plots of throughput and residual energy for the first (two degenerate) $|m| = 2$ modes.

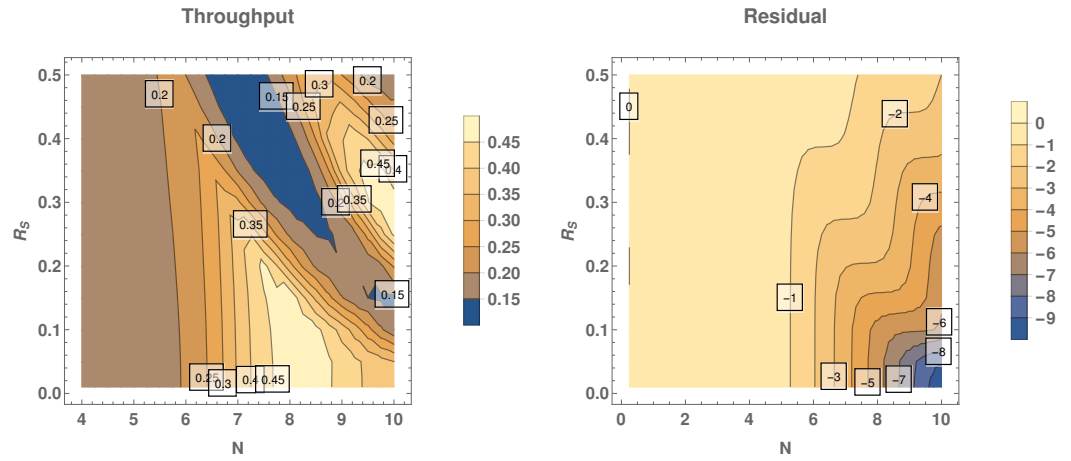


Figure 5.9: Contour plots of throughput and residual energy for the second (two degenerate) $|m| = 2$ modes.

We presently have no good explanation for this phenomenon, though we remind the reader that the ledges are only possible due to the presence of the central obstruction (Soummer et al., 2011). We believe that this is a behavior specific to the pupil geometry, as the relation appears to depend on the (m,k) labeling of the modes. Behavior very similar to this will be observed when a different pupil geometry is "close" to this, following the general rules of

perturbation theory.

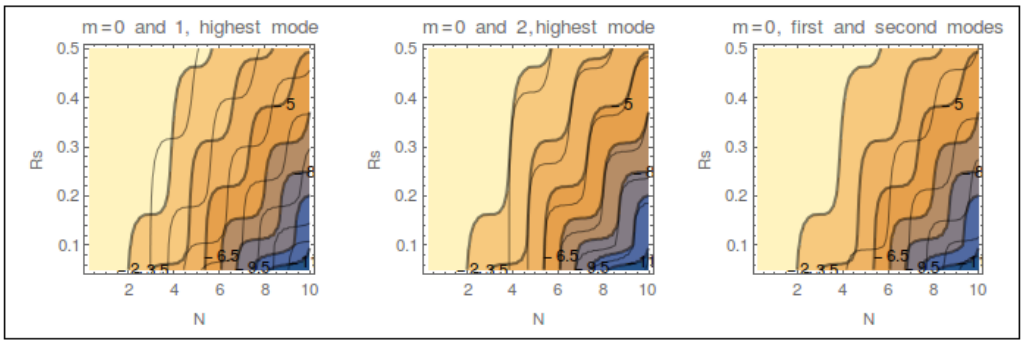


Figure 5.10: Overlay of the contour lines for the residual energies of several modes.

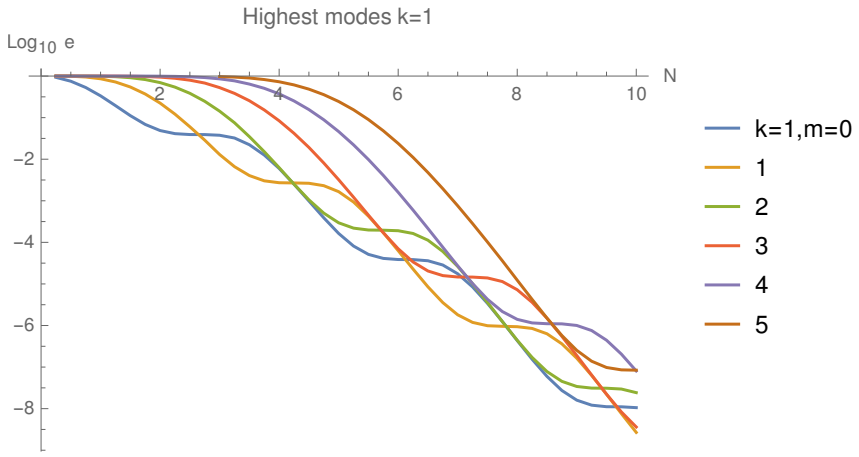


Figure 5.11: Cross-section of residual energy at $R_S = .25$ for several $m, k = 1$. $m = 5$ is not present below $\mathcal{N} = 3$ due to truncation of basis.

In (Slepian, 1964) it was shown that the eigenvalues of the CPSWFs approach a schematic form $\lambda_i \sim \left(1 + Ae^{Bi}\right)^{-1}$ for large values of the $c-$parameter.

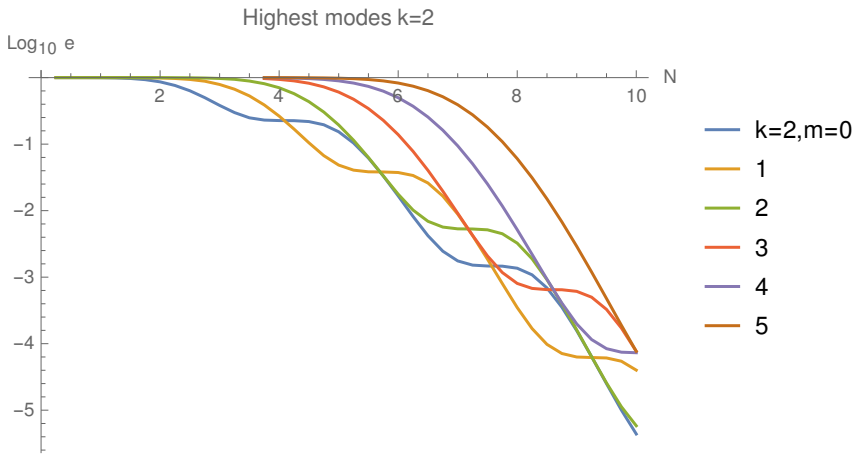


Figure 5.12: Cross-section of residual energy at $R_S = .25$ for several m, $k = 2$. Lines close to zero may be omitted due to truncation of basis.

If we consider our obstructed geometry to be the difference of two such functions of different c, then we expect the eigenvalues to follow a similar distribution. Figure 5.13 shows that we indeed observe roughly this distribution, and experimentation has shown that

$$\lambda_a = \frac{1}{1 + e^{c\left[\frac{a - \mathrm{Tr}(K)}{\mathrm{Tr}(K)}\right]}} \tag{5.3}$$

for

$$c = \mathcal{N}(1 - R_S) \tag{5.4}$$

is an eyeball-level acceptable fit, though it is not correct when fitting the data to functions of this form.

While we do not have a compelling argument for the value of the width, we do note that it is proportional to the Shannon number for the *one*-dimensional radial Slepian problem which our kernel factorization has produced. As such, we have no good prediction for the value it will take in complex geometries.

We speculated that it may be proportional to the boundaries of the regions used in the Slepian problem, but did not find this in good agreement with the JWST example of § 5.2.

5.1.4 Bell-bagel transition

The "bell-bagel" transition noted in (Soummer et al., 2009) places constraints on the desirable parameter space for APLC design, as the bagel regime is desirable given the placement of the higher-throughput regime. As a similar transition occurs around support structures, we desire a more quantitative explanation for the transition.

We know that this transition must be limited to the $m = 0$ modes, as they are the only ones which contain the necessary function $|0,0\rangle$ nonzero at the origin. Moreover, if we examine figure 5.14, we can see that the transition threshold is clearly marked by $\frac{\partial^2 \phi}{\partial r^2} = 0$, which gives a "flat-top" appearance. We may expand $\text{Œ}(r) \approx \phi(0) + \frac{1}{2} r^2 \phi^{(2)}(0) + \frac{1}{4!} r^4 \phi^{(4)}(0) + \ldots$ The location of the minima of this equation are given by $r^* = 0, \pm \frac{\sqrt{-(2/3)\phi^{(2)}\phi^{(4)}}}{\phi^{(4)}/3!}$ which can develop new minima after the above-stated threshold is crossed. With the

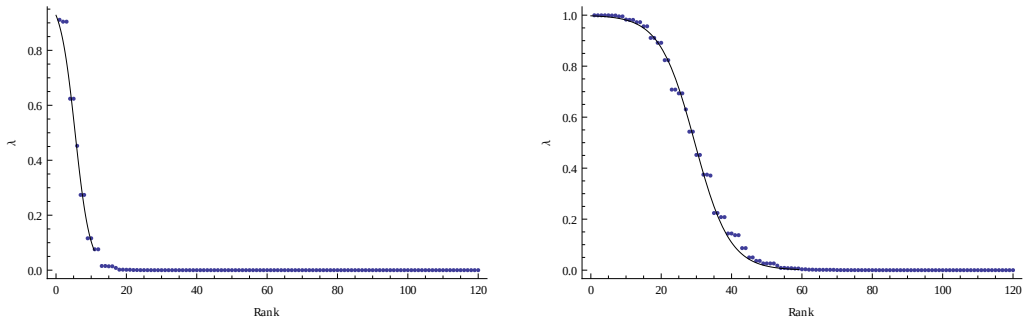


Figure 5.13: Plot of ordered eigenvalues for $R_S = 0.15$, $\mathcal{N} = 3$ (left) and 7 (right). Solid line is formula listed above, drawn to twice the Shannon number.

eigenvector coefficients for $b_{t,0}(r)$ being $V_{a;t,0}$, the transition is marked by the condition $\frac{V_{a;0,0}}{V_{a;2,0}} = \frac{1}{\sqrt{3}}$.

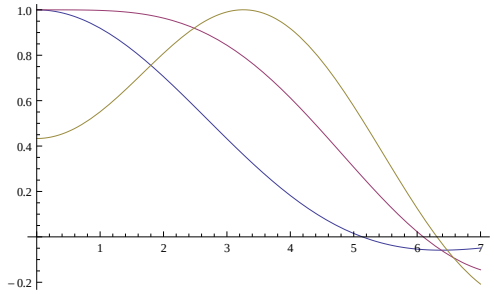


Figure 5.14: Schematic examples of bell, transition, and bagel cross-sections.

We have confirmed this condition for the highest $m = 0$ modes, as shown in figure 5.15. We have also noted the existence of "secondary" transitions, where another minimum develops inside R_S. Once again, the critical ratio aligns with the manually verified points. A third transition occurs within our parameter space, but we have not bothered to precisely locate the parameters for comparison.

Similar examination of the second $m = 0$ mode also displayed a bell-bagel transition located precisely on the $1/\sqrt{3}$ contour line, as expected.

We are not sure that this reasoning will be of particular use for asymmetric pupils, both because it relied on special properties of the $m = 0$ apodizations, and because it postdicts the result from the eigenvectors themselves instead of the parameters. Because eigenvector components are notoriously complicated functions of the matrix elements, which in this case are themselves functions of the parameters of interest, there is no simple relation to determine the location of the transition in terms of $\mathcal{N}$ and R_S. We speculate that these contours may roughly follow the ratio of the kernel elements $K_{(0,0),(0,0)} / K_{(2,2),(2,2)}$.

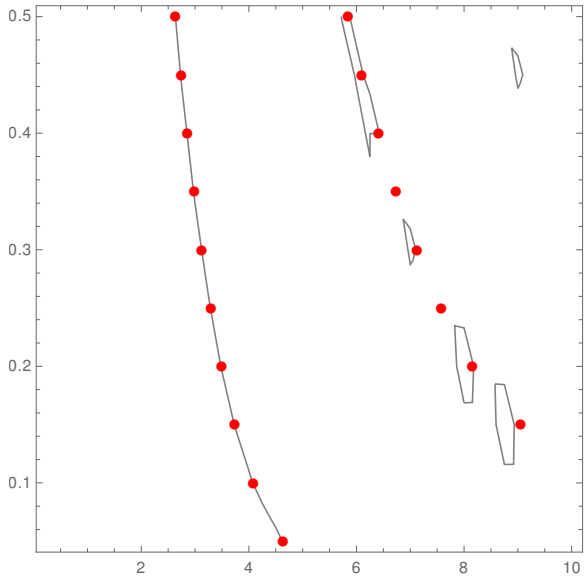


Figure 5.15: Critical contour for the eigenvector component ratio (lines) and manually checked transition points (dots). The second set of dots is the second transition described in the text. Contour lines are irregular due to interplay of data gridding and contour algorithm.

We can give a somewhat qualitative explanation for localization in angular directions. For smaller mask sizes, the number of useful radial basis modes (§ 3.3.1) is more limited. Physically, this corresponds to the wide spread of light in our reversed problem. To develop angular details of order $2\pi/\ell$ require angular modes of order $m = \ell$ and above, which only appear once the require radial modes $t \geq \ell$ become important. Once $t \geq 2\ell$ also becomes a viable basis function, then the Fourier series will become considerably more well-defined. An illustration of the shifting relevance by t and $\mathcal{N}$ is shown in figure 5.16.

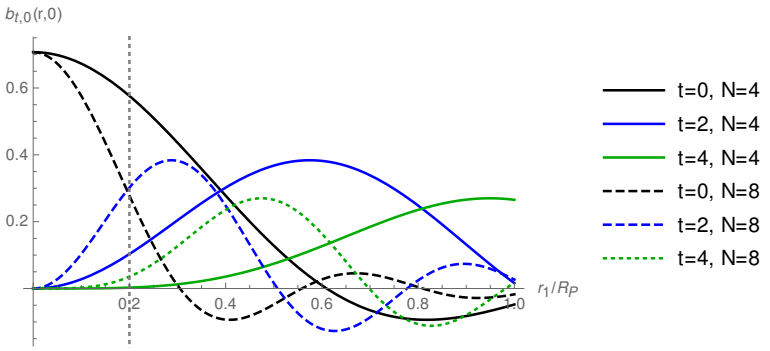


Figure 5.16: Changes in the basis functions as t and $\mathcal{N}$ shift. Gray dotted line is an example secondary cutoff; figure is scaled so that $r = 1$ is the edge of the pupil.

5.2 Non-Circular Demonstration: JWST as APLC

5.2.1 Coronagraphic set-up

The JWST (*James Webb Space Telescope*) has a 6.5m primary mirror, divided into 18 hexagonal segments. Support structures for elements of the optical system overlay the primary mirror in three places; a representation is in figure 5.17. The Near-Infrared Camera (NIRCam) (*James Webb Space Telescope*), one of several instruments, is designed for the $0.6 - 2.3$ and $2.4 - 5.0\mu$m ranges (*Near-Infrared Camera (NIRCam)*) which will be the most valuable (as discussed in 1). As a reminder, Earth at ten parsecs is a maximum of 100 mas away from the Sun, while Jupiter is 5200 mas. At one micron wavelength with this pupil, these are 3.15 and 164 λ/D_p. The camera's resolution is 31 mas/pixel (*Near-Infrared Camera (NIRCam)*), 0.978 λ/D_p at one micron.

We will examine mask sizes of widths $(D_M/L)/(\lambda/D_p) = \mathcal{N} = 2.0 - 6.0$ corresponding to wavelengths $2.50 - .833\mu$m, for $\lambda/D_p = 79.3 - 26.4$ mas. Each detector pixel is thus $0.391 - 1.17\lambda/D_p$. We will examine primarily the case where the Lyot stop is identical in shape to the pupil.

While we have been able to give the apodization in terms of a sum over basis functions in the pupil and Lyot planes, we will revert to a discrete pixelization for the instrument (D) plane. We do so by pixellizing the Lyot plane using the same rectangular mesh as was originally provided for the description of the pupil plane, which we take to be a W by W square.

It is worthwhile to here make a note of the limitation that this will apply to our D-plane calculations. Such a transform goes roughly as $\sum_j \Phi_c(j) e^{i\pi jk/W}$,

with j and k representing indices in one direction in the Lyot and instrument plane respectively. This must correspond to the continuous integral $\int \mathrm{d}^2 r\, \Omega_L(\mathbf{r})\Phi_c(\mathbf{r})e^{i\mathbf{r}\cdot\rho}$.

Recall that our possible values of r in the pupil run from 0 to $\mathcal{N}\pi/2$ even before truncation at the Lyot outer stop, so that our image covers the distance of $\mathcal{N}\pi$. With W pixels in the discretization of this plane, and ignoring phase changes associated with shifting the center, this means that $r \approx \mathcal{N}\pi j/W$, and so $ir\rho \approx i\pi\mathcal{N}\rho j/W$, which must be nearly equal to the discrete phase $i\pi jk/W$. As such, $k = \mathcal{N}\rho$ or, switching from ρ (mask radii) to ζ (distance in λ/D_P), $\zeta_k = k/2$.

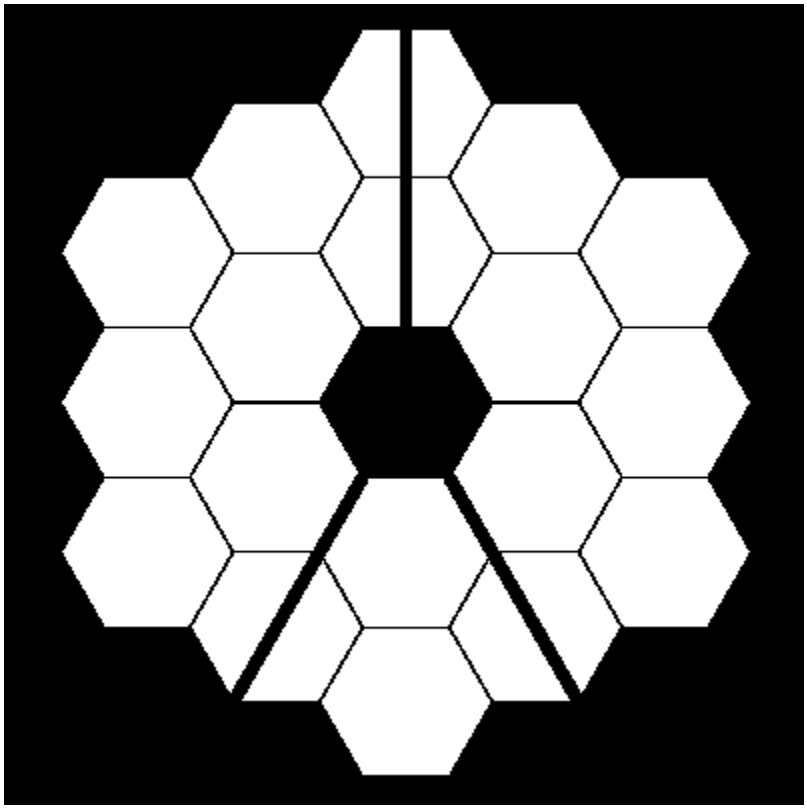


Figure 5.17: Binary representation of the JWST main mirror, 402 × 402 px.

This means that the discrete-case resolution is limited to spacings of $\frac{1}{2}\lambda / D_P$. While this is not quite fine enough to accurately model the highest-λ case, it is sufficiently close for the analysis at hand.

We created our apodizations at a design wavelength of one micron ($\mathcal{N} = 5.0$). This was done by calculating the kernel elements through (3.15),

$$K_{tm;t'm'} = \int \mathrm{d}^2 r \, P_A(\mathbf{r}) \, [b_{tm}(\mathbf{r})]^* \, b_{t'm'}(\mathbf{r})$$

The kernel elements were first generated up to $t_{max} = 9$, $m_{max} = 9$, producing 55 different (t, m) pairs. The accuracy goal of the numerical integration was limited to 4.

Accuracy checks

After the calculation of the kernel, we first verified that the limits in (t, m) and accuracy did not produce large errors in the calculated results. We did so by first estimating the magnitude of the kernel elements from larger (t, m), since these would constitute a perturbation from the truncated kernel used. The second check was comparing the sum of the eigenvalues $\sum \Lambda_a$ to the scaled area of the pupil $4\pi|\Omega_1|$, as the two should be equal by Lidskii's theorem, equation (2.6), as discussed in the end of § 3.4.

For the kernel elements, we examined the kernel value $K_{(0,0);(8,0)}$. This matrix entry was chosen as it should be the largest of the highest-t values, as it could match $m = 0$ and less affected by alternation of sign in the integrand. The kernel value was -0.00736.

The integrand evaluated at the edge of the pupil, $(b_{8,0})^*(r = \mathcal{N}\pi/2)b_{0,0}(r =$

$\mathcal{N}\pi/2)$, takes the value 0.0242, roughly three times the kernel element itself. Following the discussion in § 3.3.1, this will be the maximum value of the integrand and is, indeed, an overestimate of the kernel element.

If we assume that this roughly the same proportion as would hold for $K_{(0.0),(10,0)}$, then the prediction for that value is ± 0.0014. Neglecting this element would produce errors of this magnitude, presumably in the smallest eigenvalues and their eigenfunctions. This is less than 0.5% of the highest kernel element, and so we consider it an acceptable change for this demonstration. (The actual calculation of the kernel element yields -0.000721, below even the estimate and so justifying our neglect further.)

A plot of the kernel elements $-\log_{10} K_{00,tm}$ is shown in 5.18. The are ordered by index as $(0,0), (1,-1), (1,1), (2,-2), \ldots$ up to $(9,-9)$. We can see that the general trend is a roughly exponential decay.

The second comparison was far simpler. The relative error between the scaled area (the true Shannon number, see § (2.6)) and the sum of the eigenvalues was 7.54696×10^{-5}, acceptable for our purposes. This is reinforced by the fact that this sum included all eigenvalues, including those down sizes comparable to the discrepancy – which were the 45th in order, and so more susceptible to error than the higher-Λ_a modes of interest to us. The eigenvalues themselves are plotted in 5.19, where the exponential decay can be clearly seen.

5.2.2 Apodization and PSF results

The pupil and its point spread function (on and off axis), contrasted to the peak of the off-axis maximum, are shown in figure 5.20. The top sixteen apodization

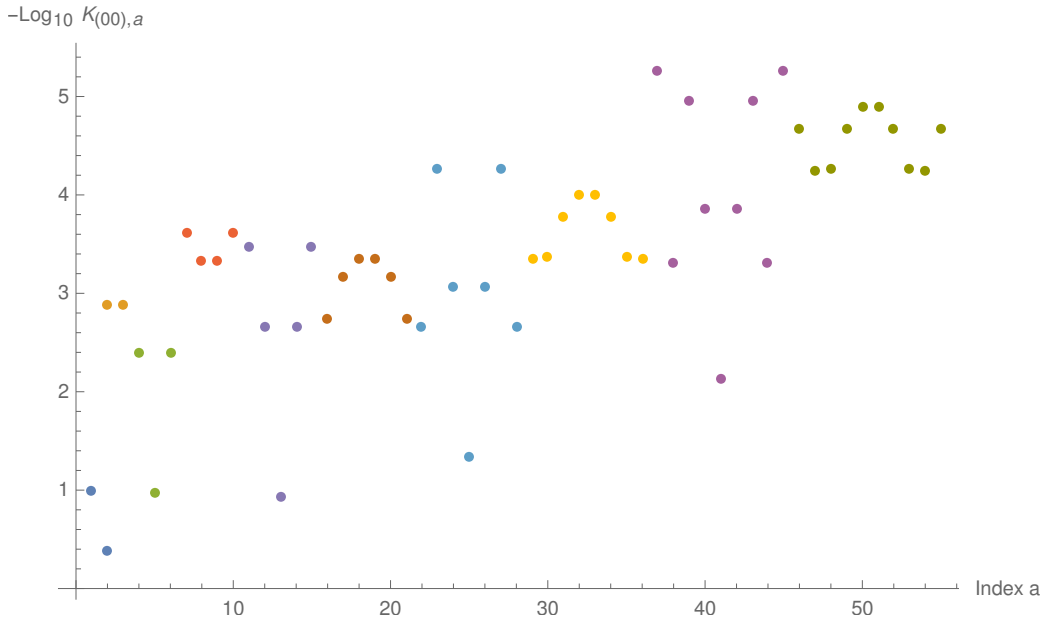


Figure 5.18: Kernel entries $-\log_{10} K_{00,tm}$ for the JWST pupil at $\mathcal{N} = 5.0$. Symmetries are $\pm m$, from use of complex $e^{im\theta}$ convention. Different colors correspond to different t.

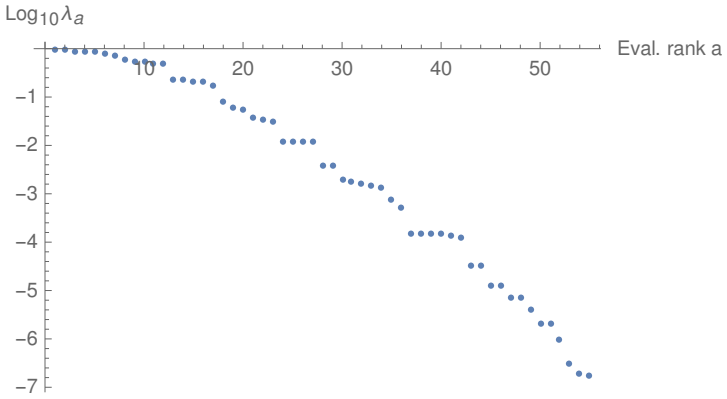


Figure 5.19: Eigenvalues of the JWST pupil at $\mathcal{N} = 5.0$.

modes are shown in figures 5.21 and 5.22; the point spread functions are shown relative to the peak off-axis PSF for that apodization.

Comparing the apodizations to the circular mask modes 5.3, even at different $\mathcal{N} = 3.5$, we can see that the eigenfunctions become highly similar by the sixth or seventh apodization, resembling trefoil and coma aberrations.

In contrast to the circular case, the highest Λ_a mode is *not* wholly positive; in fact, no mode is now wholly positive. If we desire a grayscale apodization, to avoid the difficulties associated with phase-shifting, we will need a sum of different modes.

There is also no longer a circularly symmetric mode, though $a = 10$ comes close. Such was only possible in the circular case because the angular integral in the kernel (3.15) reduced to a Kronecker delta $\delta_{mm'}$ and so separated the kernel into different blocks by angular mode, each producing its own set of apodizations. Here, the pupil mixes all different angular modes, and so there is no guarantee that $m = 0$ becomes isolated.

The point spread functions for the different modes were averaged over thin rings; that is, at all angles and radial bins $\frac{1}{2}\lambda/D_p$ wide, centered at whole and half values. They were then divided by their peak off-axis maximum value, forming the usual contrast. An example of the resulting reduction for the first apodization is shown in figure 5.23. The resulting radial contrast profiles are shown in figure 5.24. No mode matches the desired 10^{-10} contrast criterion.

The eigenvalues, maximum value in the pupil under $\langle a|a\rangle = 1$ normalization, throughput (power through apodized pupil/power through unapodized

pupil), and Lyot plane residual power (relative to that through the pupil) for the top seventeen modes are listed in table 5.1. These modes are all of those with eigenvalues above 0.1. The Lyot-plane residual power is lowered considerably from the case where the Lyot stop includes regions blocked in the pupil plane.

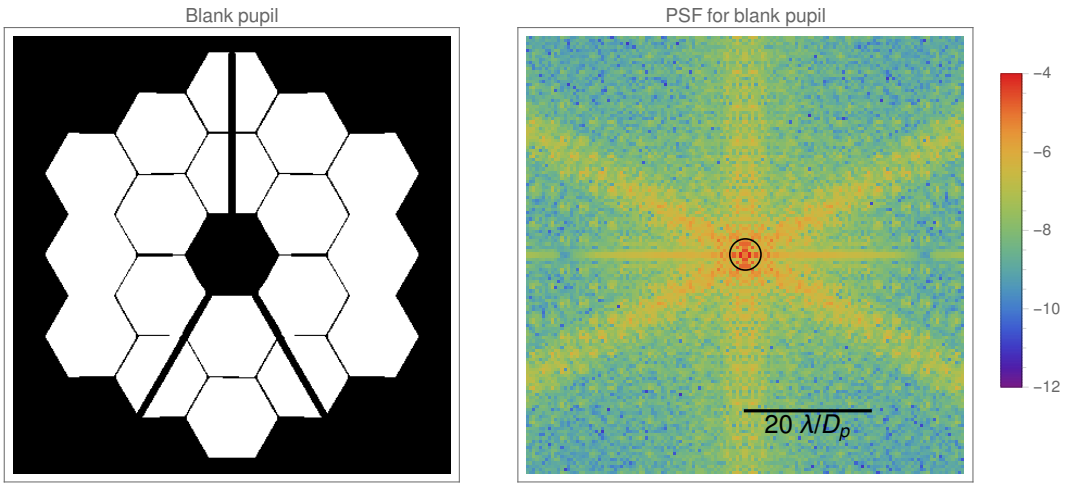


Figure 5.20: The pupil (left) and the $\log_{10}$ of the point spread function, contrasted to the off-axis maximum (right). The center circle is the projection of the mask in the latter.

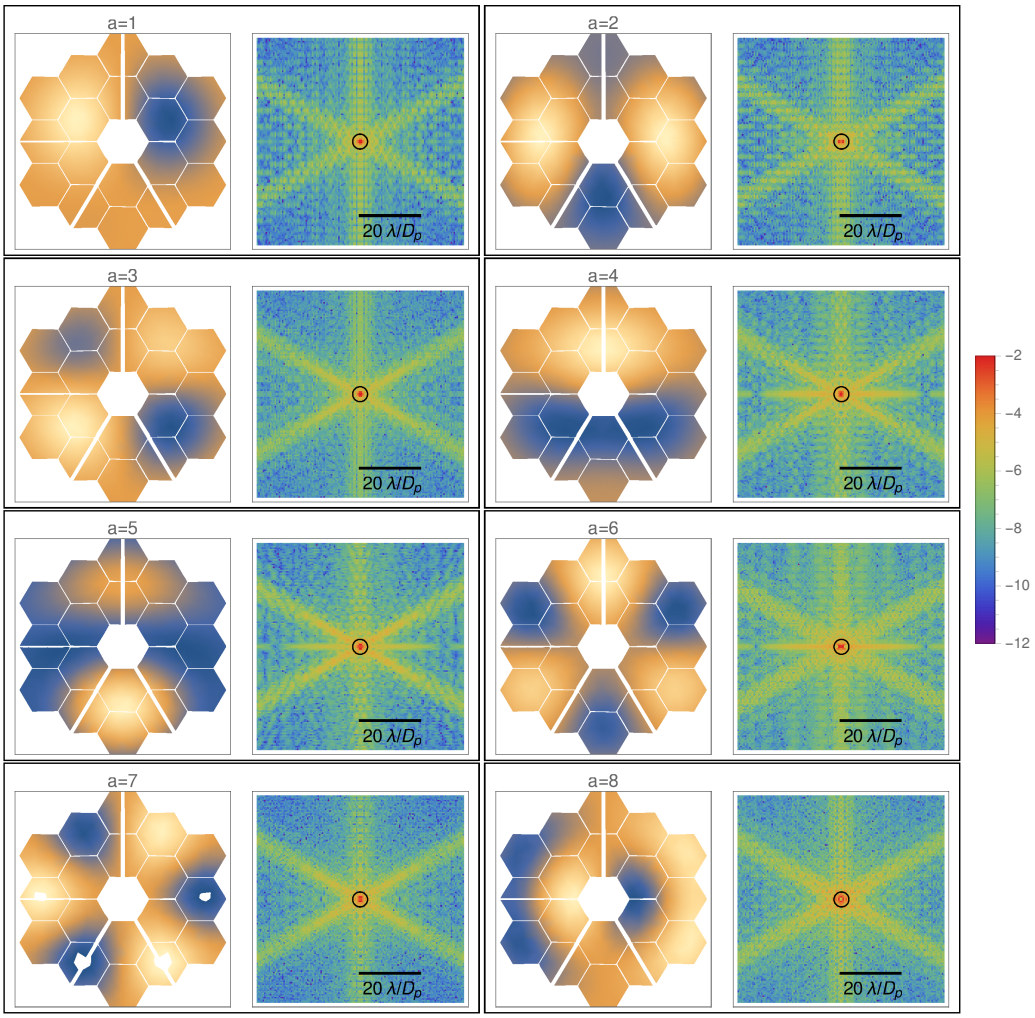


Figure 5.21: Left half of pairs: the top eight Slepian mode eigenfunctions of the JWST pupil at $\mathcal{N} = 5.0$. Right half: the $\log_1 0$ PSF's of those apodizations relative to their peak off-axis intensity.

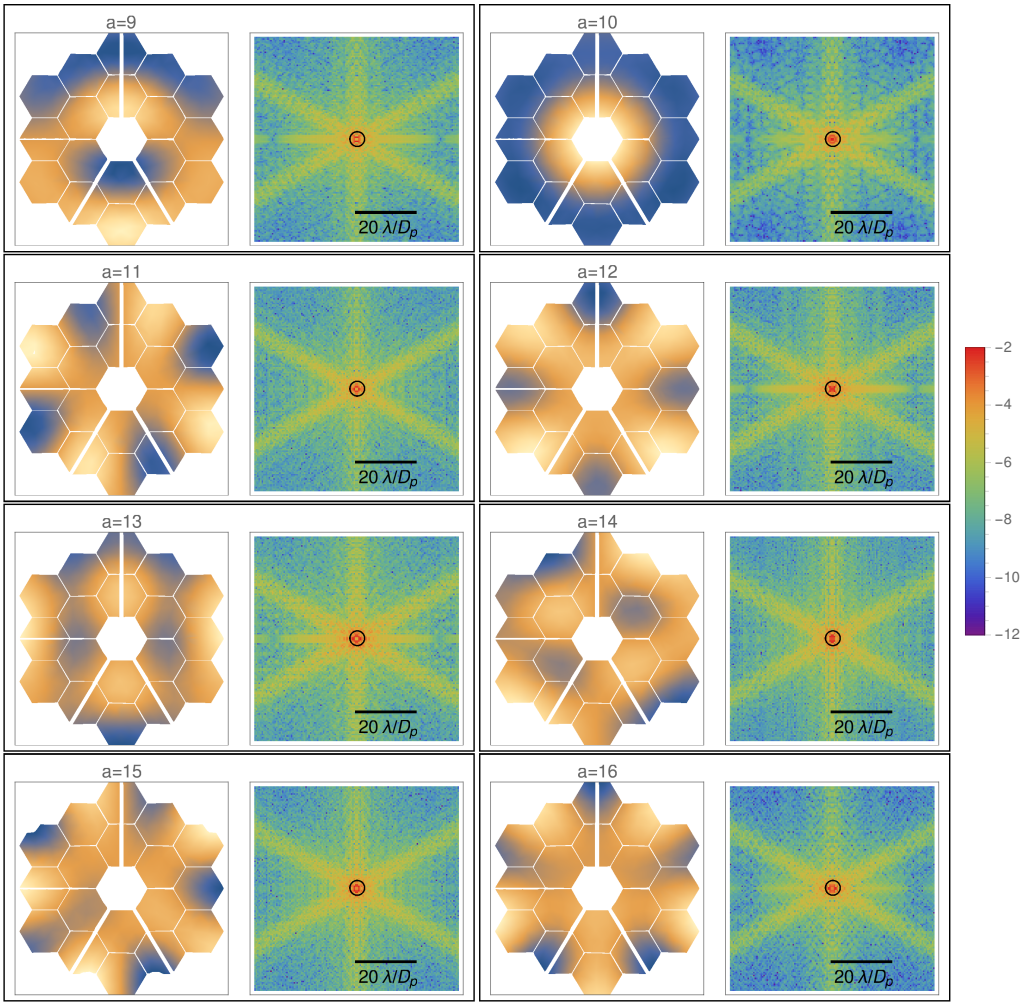


Figure 5.22: Left half of pairs: the ninth through sixteenth Slepian mode eigenfunctions of the JWST pupil at $\mathcal{N} = 5.0$. Right half: the $\log_1 0$ PSF's of those apodizations relative to their peak off-axis intensity.

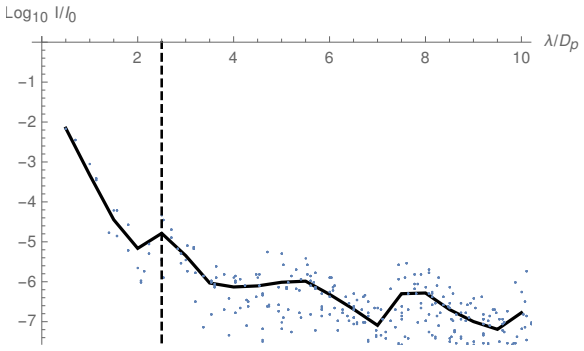


Figure 5.23: Example of the resulting averaging reduction process to produce the first mode's radial contrast profile.

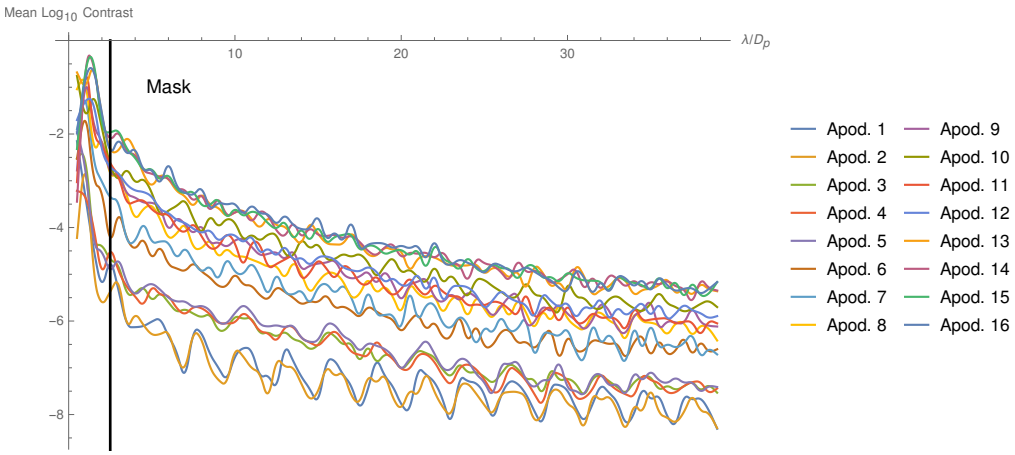


Figure 5.24: Averaged radial contrast profiles for the top sixteen apodization modes.

Rank a	Λ_a	$\mathcal{F}$	throughput	Lyot residual power
1	0.916304	0.1868	0.2112	0.007005
2	0.909764	0.1807	0.2240	0.008141
3	0.882031	0.1825	0.2131	0.01391
4	0.868486	0.1860	0.2019	0.01730
5	0.856846	0.2368	0.1229	0.02049
6	0.77102	0.1785	0.1947	0.05243
7	0.689809	0.1302	0.3272	0.09622
8	0.574585	0.1268	0.2876	0.1810
9	0.550885	0.1545	0.1857	0.2017
10	0.521777	0.1973	0.1078	0.2287
11	0.477616	0.1249	0.2462	0.2728
12	0.473909	0.1596	0.1497	0.2767
13	0.226115	0.1263	0.1140	0.5988
14	0.226053	0.1228	0.1206	0.5989
15	0.210886	0.1064	0.1500	0.6227
16	0.202405	0.1116	0.1307	0.6361
17	0.166358	0.1103	0.1101	0.6948

Table 5.1: Eigenvalue, maximum value in pupil, normalized throughput, and Lyot stop power relative to pupil transmitted power, in the case where the Lyot stop is equal to the initial pupil.

5.2.3 Instrument Plane Response to Combined Apodizations

It is clear that no single apodization has the features that we desire, and we must combine apodizations. To ensure that our starting point is positive, we can construct the highest $m = 0$ mode for a circular aperture of the same pupil radius. We can further estimate that this circular mode has an effective circular secondary obstruction interior to the hexagonal hole at the center, with estimated $R_S \approx 0.15$. We have established in the beginning of § 5.1.2 that this must be a positive function in the circular pupil.

Since we have constructed this imaginary pupil to wholly include the physical one, this function will also be wholly positive inside our arbitrary pupil. As this is then a function defined inside the pupil, we can decompose it in terms of the eigenfunctions. Denoting this circular mode as U_{tm}, its overlap with an eigenvector in the pupil is $\langle a| P_1 |circ\rangle = \Lambda_a (V^a)^*_{tm} U_{tm}$.

Unsurprisingly, very few modes overlap with the circular mode. Most of the higher eigenfunctions are effectively purely higher-m modes, and so have negligible overlap. The top three overlapping eigenfunctions for the pupil at hand are $(2, 5, 6)$, with overlap coefficients $(0.622383, 0.591839, 0.100872)$. Together, these three modes account for 99.4% of the power coming through the circular function inside this pupil.

Figure 5.25 shows the apodization produced by summing those three modes, weighted as the coefficients demand, and its PSF (contrasted to its off-axis maximum). A radial profile, averaged as with the profiles in figure 5.24, is shown in figure 5.26.

The contrast comparable to that of apodization #2, but still several orders of magnitude above the desired level. The residual energy (in the Lyot stop shaped like the pupil, relative to that through the pupil) for the blank pupil is 0.0502, while that for the circular mode is 0.0151. For the three modes

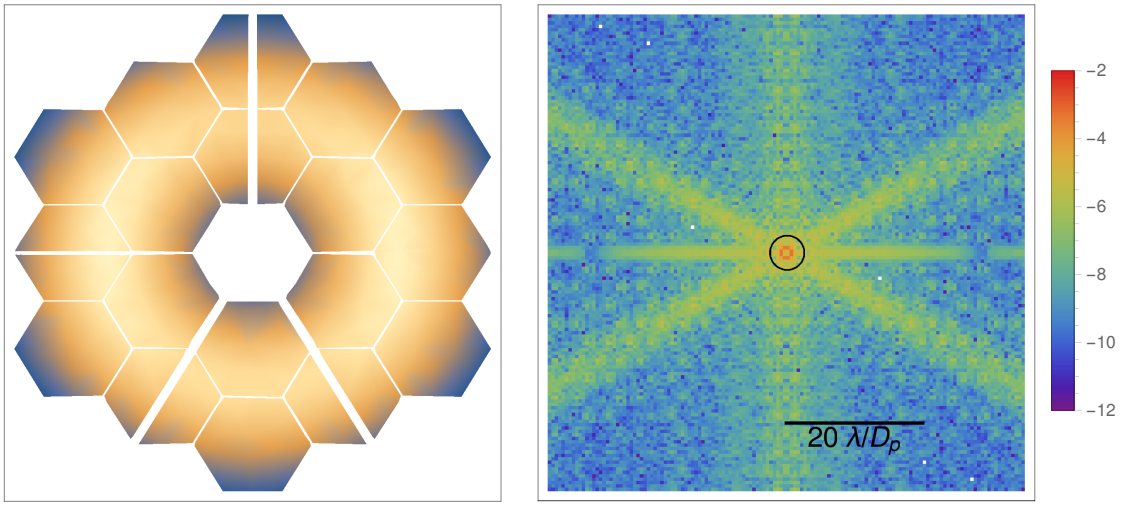


Figure 5.25: Circular, positive apodization produced by combining modes 2, 4, and 6 as determined by overlap coefficients.

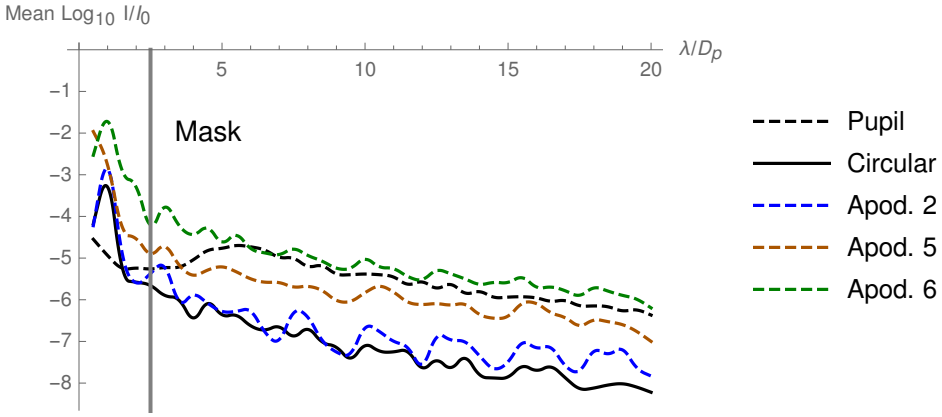


Figure 5.26: Circular positive apodization-produced PSF contrast, averaged as in figure 5.24. The three modes used to build the circular mode, and the blank pupil, are shown for contrast.

involved, we had values 5.1 0.00814, 0.0205, and 0.0524.

The predicted value in this case, $\sum_a \frac{|\alpha_a|^2 \Lambda_a}{\sum_b |\alpha_b|^2 \lambda_b} \times resid_a$ (§ 4.1.2), yields 0.0143, showing that the very crude pixellization process employed has a relative 5.5% error.

The throughput (power transmission relative to a blank pupil) is a respectable 0.5445 for this circular case, contrasting with values of 5.1 0.224, 0.123, and 0.195 for the individual modes.

Several facts stand out. First, the blank pupil has a lower relative residual power than many of the apodization modes, beginning as high as $\Lambda_a \approx 0.77$. Second, and related to the first, is that the contrast for the blank pupil is better than many of the individual modes. Third is the fact that the throughput for the circular mode is considerably better than that of any of the individual modes, even those that were not used to build the circular apodization.

Since the circular apodization has failed to reach acceptable contrast, we must clearly explore variations away from it that retain positivity. A first check is using all of the overlap coefficients, rather than just the first three; this produces a calculated residual of 0.0150 (compared to the prediction of 0.0149), and a throughput of 0.5467, and slightly worse on the contrast levels shown in figure 5.27.

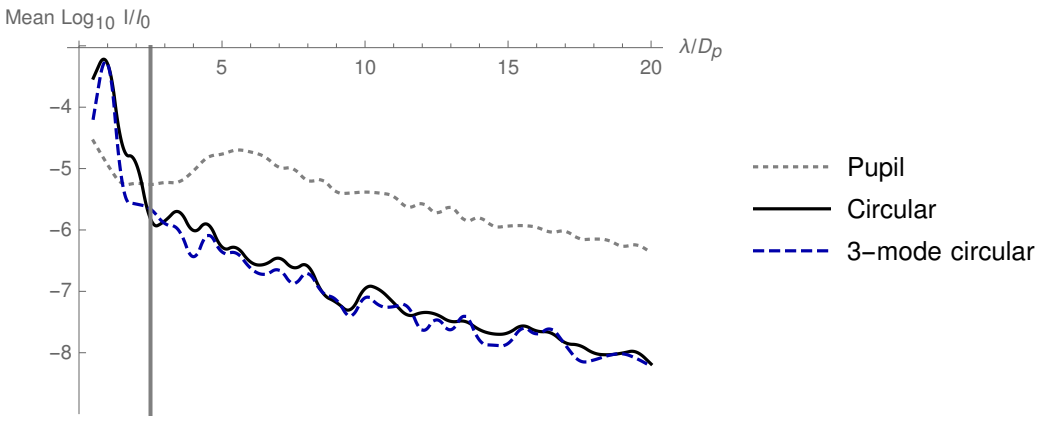


Figure 5.27: Circular positive apodization-produced PSF contrast, averaged as in figure 5.24, using the top 17 modes' overlap coefficients. The blank pupil and the 3-mode circular apodization are shown for comparison.

5.2.4 Eigensystem Trends with Changing Wavelength

Now that we established the behavior of the coronagraph at that specific $\mathcal{N}_0 = 5.0$, we wish to understand a little behavior across a wide band of wavelengths. We will study mask sizes from $\mathcal{N} = 2$ to 6, corresponding to wavelengths of 2.5 to 0.833 microns.

We begin by using the dilation method, § 4.2, specifically formula (4.24), to calculate the kernels for the other desired wavelengths:

$$K_\eta = \eta^2 \, D_\eta K D_\eta^T$$

where $\eta = \mathcal{N}_{new}/\mathcal{N}_{old} = \lambda_{old}/\lambda_{new}$ ranges from 0.4 to 1.2.

Broadband Numerical Stability

We first consider whether the dilation method is sufficiently stable for complicated kernels. Our concern focuses around η far from unity, both above and below. We must firstly check if we have included enough basis modes to allow dilation usage, as we have argued in 4.2.3 that convergence will need more modes than we normally choose to behave at all.

Even with sufficient inclusion of basis functions, we may still have unacceptable errors. By limiting the accuracy of the numerical integration in the kernel, dilations significantly different from the identity will accumulate increasing error, and therefore deviate from the true kernel.

To address these concerns, we first use the dilation method on the kernel derived above using only up to $t = 9$. The relative error in the true (area$/4\pi$) and estimated ($\sum \Lambda_a$) Shannon numbers for the spread of $\mathcal{N}$ are shown in

figure 5.28.

We can see that all of the $\eta > 1$ calculations are untrustworthy; the error on $\eta = 1.2$ was 25%! Relative error down to $\eta = .70$ was 10^{-4} or less, though it began to increase exponentially again below this value.

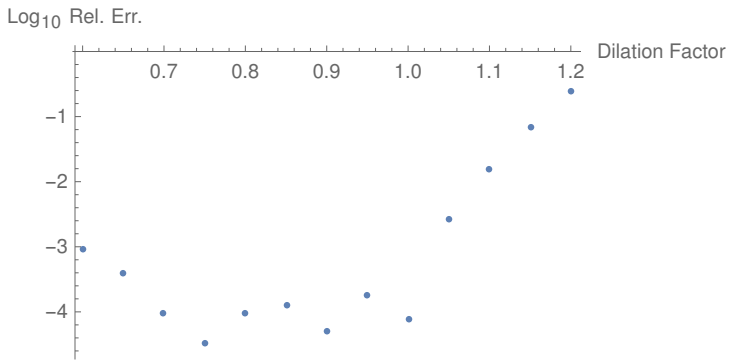


Figure 5.28: Log10 of the relative error in the Shannon numbers for a variety of dilation factors, using up to $t = 9$.

To determine the source of the errors, we repeated the $\mathcal{N} = 5.0$ kernel calculation, but this time included up to $t = 15$ modes. The accuracy goal of the numerical integration remained 4, unchanged from the $t = 9$ calculations. The relative errors are shown in 5.29.

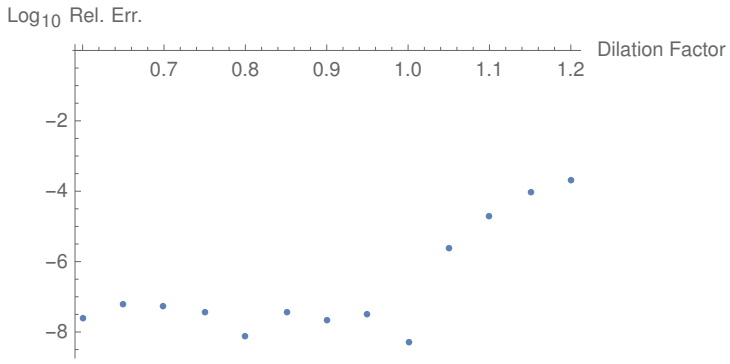


Figure 5.29: Log10 of the relative error in the Shannon numbers for a variety of dilation factors, using up to $t = 15$.

With these basis modes included, we see that the $\eta < 1$ calculations have the same level of relative error as the original $\mathcal{N} = 5.0$ kernel. The $\eta > 1$ calculations still show a large growth in the relative error, though its magnitude is considerably decreased from the $t = 9$ case. It would seem that even with a large inclusion of basis functions, the error in the kernel elements continues to compound itself in an exponential fashion with increasing η, though this can be reduced to acceptable levels with sufficient accuracy in the numerical integration.

Given these results, we conclude that it will be safest to always select the largest $\mathcal{N}$ (minimal λ) of interest in a bandwidth and calculate all other kernels of interest using $\eta < 1$. This will avoid both the concern of including insufficient basis modes, as well as preventing the growth of errors.

We will use the $t = 15$ kernels for the remainder of this section. Given the low relative error in the $\mathcal{N} = 6$ result, we will accept the results from dilating the $\mathcal{N}_0 = 5$ kernel in this case.

Trends

Now that the acceptability for the dilation-method kernels is established, we can show the resulting behavior of the eigensystems. The behavior of the top eight eigenvalues across the dilation is shown in figure 5.30, and the next eight in 5.31. (For ease of comparison to circular results, the first eight residual energies $(1 - \Lambda_a)^2$ are shown in 5.32) As expected, all increase with increasing $\mathcal{N}$, as proven to occur in 4.2.

The peculiar jumps visible come from crossing points of the curves; since

we have labeled our eigenvalues in descending order, seemingly distinct functions will switch labels. We can see such an example in figure 5.33, where the second-ranked mode at $\mathcal{N} = 2.75$ becomes the first-ranked mode at $\mathcal{N} = 3.25$. At this crossing point, it is impossible to distinguish between the modes, as any two orthogonal combinations are equivalent. We have not explored any physical significance to these crossing points.

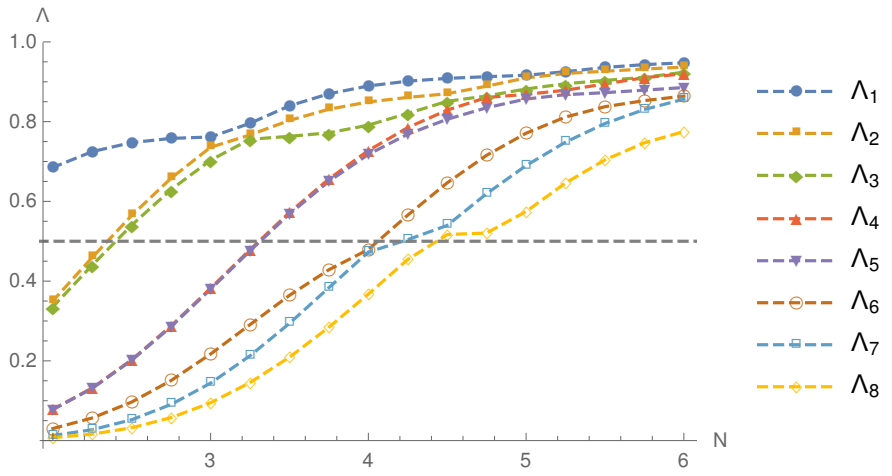


Figure 5.30: Top eight eigenvalues for the various $\mathcal{N}$.

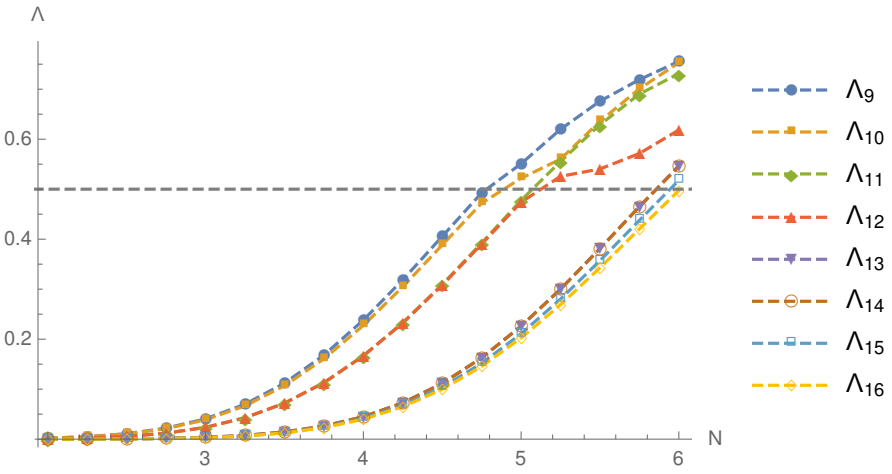


Figure 5.31: Next eight eigenvalues for the various $\mathcal{N}$.

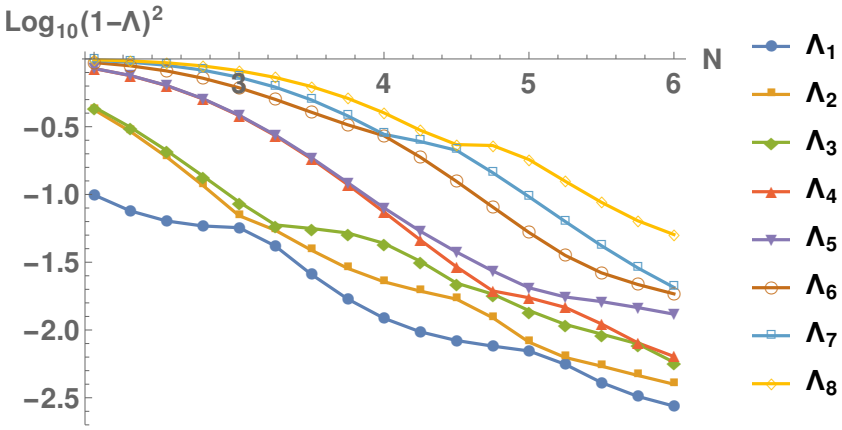


Figure 5.32: Top eight $\log_{10}(1-\Lambda_a)^2$ for JWST.

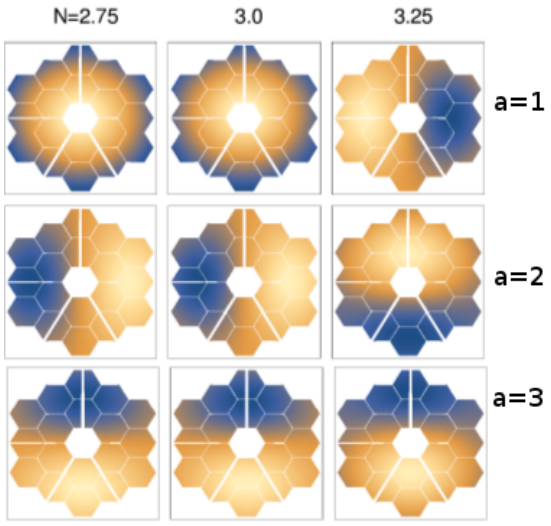


Figure 5.33: Apodizations ranked $1, 2, 3$ between $\mathcal{N} = 2.75$ and 3.25. The mode which begins as number 2 changes ordering to become number 1.

The derivatives of some of the eigenvalues are shown in figures 5.34, 5.35, and 5.36, to emphasize the achromatic regions of each. The discontinuities re-emphasize the trouble which occurs when the relative ranks of different modes switch.

When this crossing and rank-switching are accounted for, it appears that the eigenvalues are shifted and scaled versions of the same function in $\mathcal{N}$. Since this is no long a function of mode number a it cannot correspond exactly to (5.3), but appears highly similar. The shifting and scaling dependence on the geometry would be a useful future direction for research.

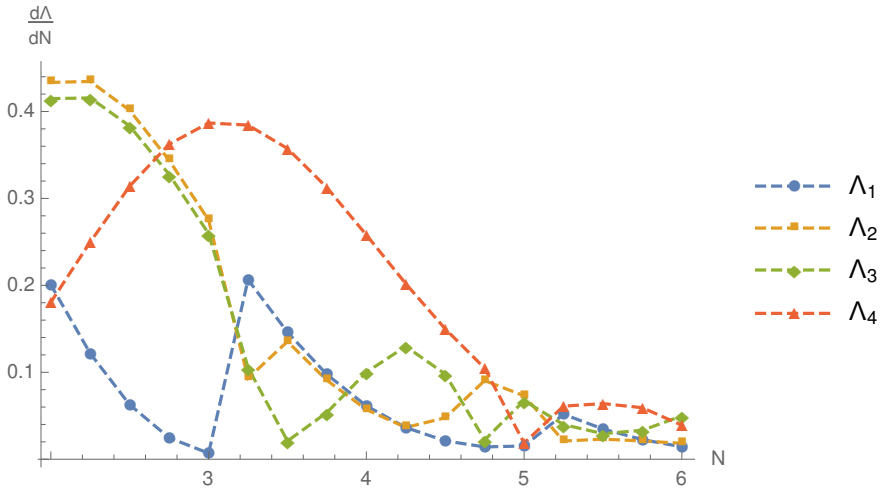


Figure 5.34: Derivatives of the first four eigenvalues for the various $\mathcal{N}$.

The first eight apodization modes, for each integral $\mathcal{N}$ in the range, are shown in figures 5.37 and 5.38. More shifts in ordering are evident. Moreover, we see that at $\mathcal{N}$ where identifiable modes shift, others seem to disappear. Consider the $\mathcal{N} = 2.0$ $a = 1$ circular mode. It is still distinct at $\mathcal{N} = 3.0$, but at $\mathcal{N} = 4.0$ disappears as a separate apodization. It seems to have combined with the initial $a = 3$ mode to produce the $\mathcal{N} = 4$ $a = 3$ mode, even while that

function alone becomes the new $a = 2$.

Likewise, the initial circular $a = 6$ mode, which seems to combine with the initial $a = 7$ trefoil at $\mathcal{N} = 4.0$ to produce the unusual triangular apodizations. Other modes remain distinctive throughout a larger range. The initial $a = 8$ trefoil, for instance, is ordered as $a = 8, 8, 8, 7, 7$ without much ambiguity. We

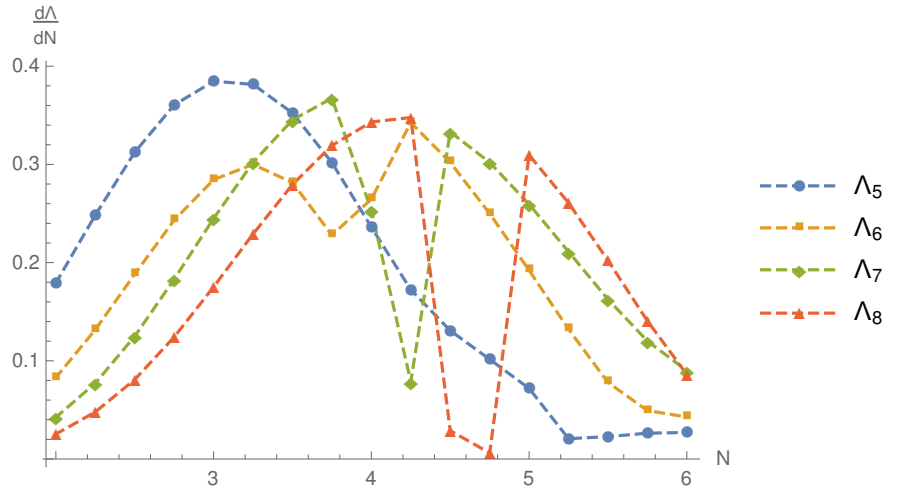


Figure 5.35: Derivatives of the eigenvalues Λ_5 to Λ_8 for the various $\mathcal{N}$.

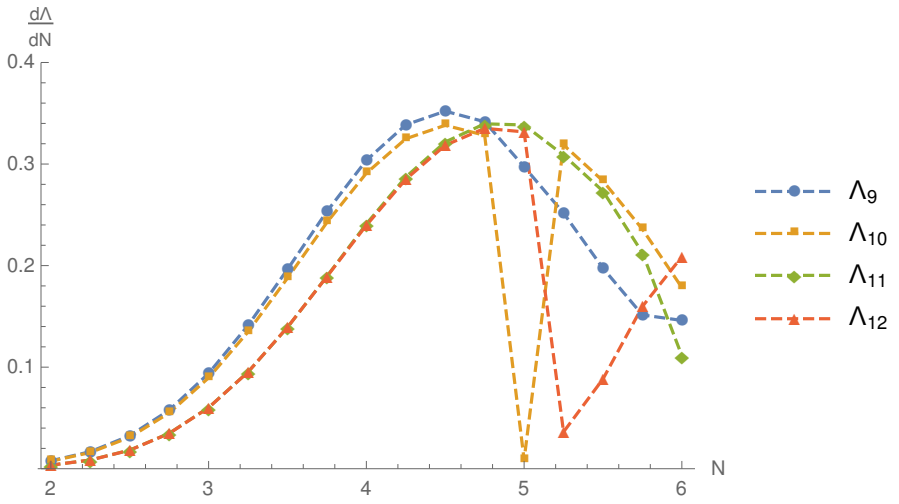


Figure 5.36: Derivatives of the eigenvalues Λ_9 to Λ_{12} for the various $\mathcal{N}$.

can identify this change of ordering in the eigenvalue plot 5.30, where the $a = 7$ and 8 mode lines seem to cross at $\Lambda = 1/2$ around $\mathcal{N} = 4.5$.

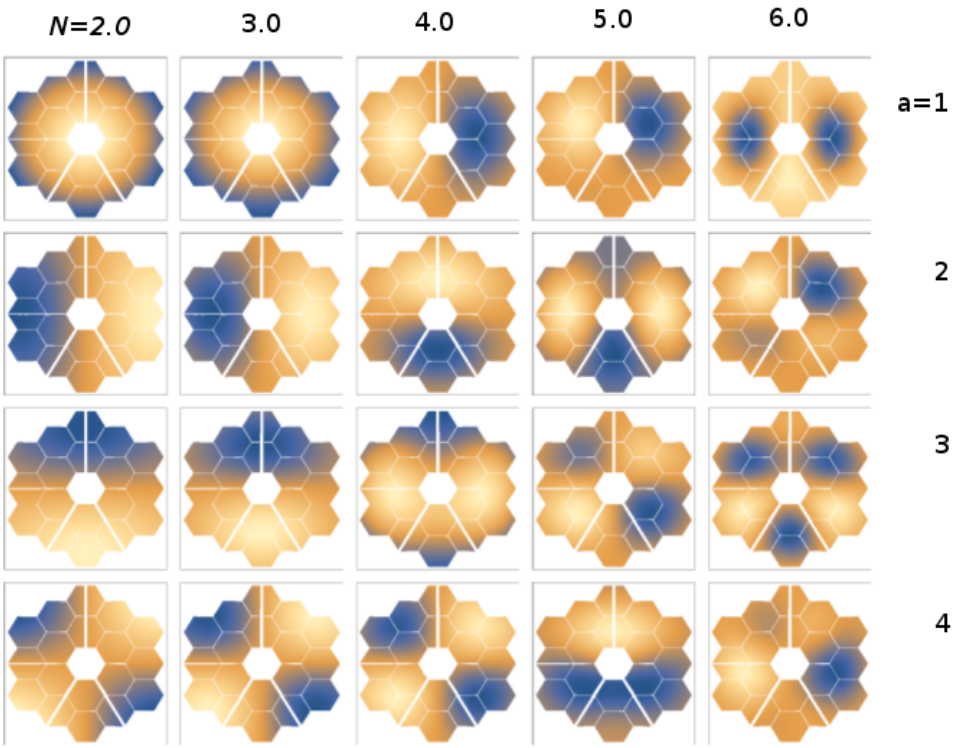


Figure 5.37: Highest-ranked four apodization modes.

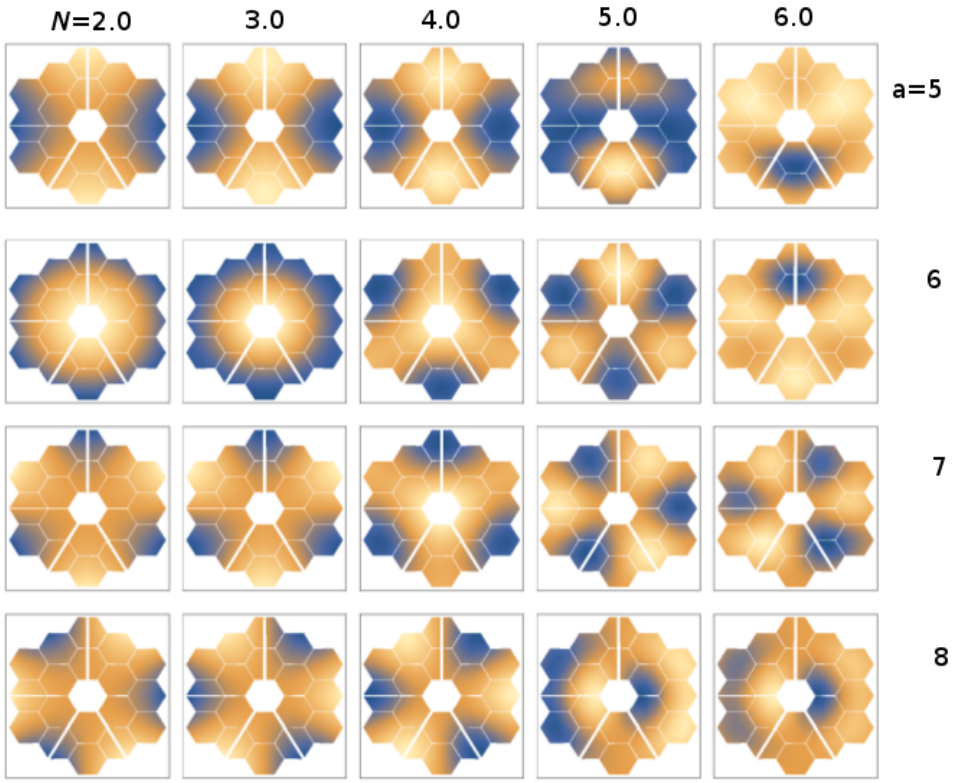


Figure 5.38: Fifth through eighth ranked apodization modes.

For a fixed $\mathcal{N}$, the distribution of eigenvalues shown in figure 5.39 is fairly similar to the circular case. The deviation from the exact form is attributed to the difference from that case.

By-hand fitting of the distributions for a few $\mathcal{N}$ to (5.3), we confirmed that $\mathrm{Tr}(K)$ scales as $\mathcal{N}^2$ as expected, and also that the width parameter $c \propto \mathcal{N}$ to excellent approximation. These are shown in 5.40. If this linearity is a universal feature, then the point $\Lambda_a = 1/2$ for any a will nearly coincide with

$$\frac{\partial^2 \Lambda_a}{\partial \mathcal{N}^2}\Big|_{\Lambda=1/2} = 0$$

for *all* eigenvalues; i.e., $\Lambda = 1/2$ is the least-stable eigenvalue.

We have no proof that this must be so, and are cautioned by the fact that this formula is not exact, but nonetheless are willing to wager it is so.

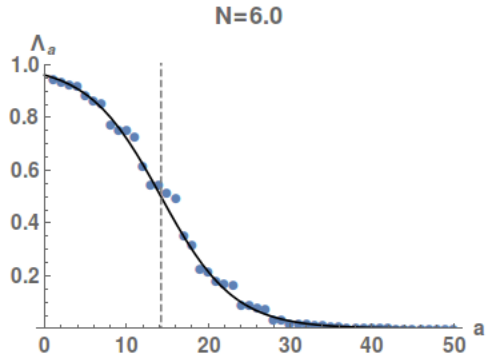


Figure 5.39: Distribution of eigenvalues in a at fixed $\mathcal{N} = 5.0$. Dotted line is predicted $\Lambda_a = 1/2$ value from the Shannon number; black line is hand-fitted distribution (5.3).

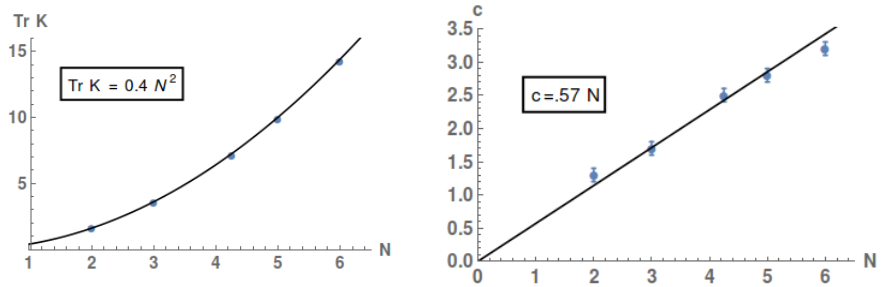


Figure 5.40: Values for the central location (left) and width (right) for hand-fitting (5.3) to the eigenvalues at different $\mathcal{N}$. Uncertainties estimated from hand-fitting procedure; black lines are approximate fits to trend.

5.3 Conclusions

5.3.1 Circular

The behavior of circular pupils with Lyot masks under our formulation matches the previously observed behavior. We are able to extend these previous results to develop the behavior of higher modes, which factorization of the kernel allows us to identify with angular and radial numbers instead of solely by eigenvalue index a. Considerable other analytical simplifications likewise occur.

The eigenvalues of the odd-angular modes appear to reach achromatic regions at the most sensitive points of the even-angular modes, and vice versa. While we do not have an explanation for this observed behavior, the precision with which it occurs makes it appear likely to us that there is a formal proof possible. We do not believe that such a proof will apply to non-circular pupils; however, if such are close to circular, then their modes will behave similarly.

For each set pupil geometry, the distribution of eigenvalues appears to roughly follow a universal formula depending on index a, mask size $\mathcal{N}$, and secondary radius R_S. This formula is in keeping with the requirements of eigenvalue change in a as described in 2.4. It is also effectively symmetric about the value $\Lambda = 1/2$ if we treat a as a continuous label instead of a discrete index. The closest values to $\Lambda = 1/2$ occur precisely when a is the Shannon number. The width of the transition from $\Lambda \approx 1$ to $\Lambda \approx 0$ is proportional to the effective one-dimensional Shannon number. (This is not the actual one-dimensional Shannon number, as the modes available are limited by the

angular mode m.) We do not have an explanation, nor a generalization to non-circular pupils, for this formula.

The noted bell-bagel transition, a qualitative shift in the appearance of the all-positive mode, is shown to be perfectly correlated with the ratio of two eigenvector elements predicted mathematically. While this is a postdiction, we have qualitatively explained this in terms of the relative amounts of energy these two basis modes contain (i.e. the kernel elements). This allows us to discuss similarly observed separations in noncircular pupils around spiders and other such secondary lines. Such require high angular basis mode number to resolve, which requires high radial basis mode number. Such only become important as the mask size $\mathcal{N}$ becomes large.

The simplifications which occur to circular pupils are very nearly matched by those where the only secondary structures other than the central obstruction are sectors (direct radial lines removing an angular width). While not all of the simplifications available in this chapter would be likewise present, we know that the kernel elements would continue to be analytical functions. The instrument plane average intensity and encircled energy expressions might enjoy a continued simplification over the fully general case even when the Lyot stop does not match the pupil but remains circular.

5.3.2 Non-Circular

The capability of the apodization algorithm to handle arbitrary apertures is well proven. Using the JWST primary mirror as a demonstration example, and a mask size $\mathcal{N} = 5.0$ showed that the apodization functions rapidly come

to resemble the familiar Zernike classes (coma, trefoil etc.), with only the first few showing the gross influence of the pupil's shape. Larger mask sizes allow more modes to be influenced and uniformly raise their eigenvalues.

Unlike in the circular case, no all-positive mode is generated. Since we wish to avoid phase-related apodization on the pupil, we need to find a combination of modes that passes both this and the contrast criterion. A good starting point for this search is the highest-eigenvalue Slepian found from the circular pupil of the same size as the complicated geometry, though in the JWST case the resulting contrast levels were unacceptably high. Useful solutions will need to take this as a starting point for further refinement.

The method of dilation, derived in 4.2 and explored for circular geometry in 5.1, continues to be valid in the JWST geometry. However, the limited number of base modes used in the kernel meant that the dilation method only converged satisfactorily for $\eta < 1$. Using this dilation not only results in highly significant gains in speed, but also allows us to discuss the behavior of apodizations at different $\mathcal{N}$ in terms of each other.

In comparing the broadband results, we discovered that the relative ordering of the eigenvalues can be said to change if we consider similar Slepian modes at different η to be a continuous change of a single mode. This is in contrast to the circular case, where the relative ordering of the even and odd parts were each independently fixed. We are not aware of physical significance attached to these transitions, but believe that investigation of such would be a worthwhile endeavor.

At the point $\Lambda = 1/2$, the eigenvalues for this geometry seemed to have

their steepest change in $\mathcal{N}$. If the mysterious c parameter from (5.3) is proportional to $\mathcal{N}$ always, not just in the circular case, then this will be a universal fact independent of geometry. While we believe so, genuine proof is required. If true, then π phase masks will always face terrible chromatic issues on top of the dispersive effects.

References

Soummer, R., L. Pueyo, A. Ferrari, C.Aime, and A. Sivaramakrishnan (2009). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures, II. Theoretical Properties and Application to Extremely Large Telescopes". In: *The Astrophysical Journal* 695.1, pp. 695–706.

NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.14 of 2016-12-21. URL: http://dlmf.nist.gov/.

Soummer, Rémi, Anand Sivaramakrishnan, Laurent Pueyo, Bruce Macintosh, and Ben R. Oppenheimer (2011). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures. III. Quasi-Achromatic Solutions". In: *ApJ* 729.144.

Slepian, D. (1964). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — IV: Extensions to Many Dimensions and Generalized Prolate Spheroidal Functions". In: *The Bell System Technical Journal* 43.6, pp. 3009–3057.

James Webb Space Telescope. URL: https://jwst.nasa.gov/.

Near-Infrared Camera (NIRCam). URL: https://jwst.stsci.edu/instrumentation/nircam.

Chapter 6

Conclusion

Our desire to directly observe the Jupiter- and Earth-like exoplanets we now know populate the galaxy has driven advancements in coronagraphic techniques. The instruments required need to reduce starlight by levels of 10^{-9} to 10^{-10} at angles below an arcsecond off-axis. A large number of possibilities have been developed, which require careful numerical simulation of the wavefront propagation in order to meet the useful scientific design criteria. Each of these require sensitivity analysis for various aberration possibilities, responsiveness to differing wavelengths, photon noise, and other such real-world effects.

It was realized in (Aime, Soummer, and Ferrari, 2002) (with extension in (Soummer, 2005)) that apodization for solid and π phase mask designs come from the spheroidal prolate functions, based on earlier work (Slepian and Pollak, 1961) (Slepian, 1964) (Slepian, 1976). This allowed both for explicit solutions to circular pupils and an algorithmic approach to converge on a solution for arbitrary pupils (Soummer et al., 2009). This solution is one of a family of eigenfunctions of a coronagraphic operator. While this algorithm,

based in part on Gram-Schmidt orthogonalization, is capable of determining other functions in the family, it develops errors too rapidly to determine more than a few.

Results

We have taken this mathematical fact and explored more of the implications, finding that this category of "Slepian" problems naturally encompasses the propagation of light in all finite-mask coronagraphy in the Fraunhoffer regime. Building on this realization, we were able to show that the optically reversed problem of illuminating a pupil from the mask would give us a finite and tractable matrix, whose eigenvalues and eigenvectors corresponded to the entire family of solutions.

This matrix is the *kernel*. It corresponds to the propagation operator for the optical reversal, from mask through the pupil and to a fictitious second image plane. Despite this, it is entirely possible to use the resulting functions to describe propagation from the pupil to the Lyot stop. Our investigation is considerably aided in adopting notation based on Dirac's bra-ket formalism of quantum mechanics, as this allows us to turn our focus more on abstract properties than specific integrals.

We have been able to show that the finite size of the matrix, in comparison to the true infinite-dimensional nature of the problem, introduces a controllable amount of error into the solutions. Moreover, this error decreases exponentially with the number of entries, as it scales with the eigenvalues which must likewise shrink.

This approach does not rely on discretization of the pupil and mask planes to produce the matrix. Instead, it uses a set of basis functions in which to expand the apodization and electric field. This is a familiar strategy; wavefront aberrations have long been described by coefficients of the Zernike polynomials over the pupil. Reflecting our optical reversal, our basis functions are Zernike polynomials over the mask, whose Fourier transforms are $1/r$ times the Bessel "J" functions. Similarly, our coordinate system is such that the mask edge is located at a value $\rho = 1$ instead of the pupil's. That radius is at $r = \mathcal{N}\pi/2$, with $\mathcal{N} = (D_M/L)/\lambda/D_P$.

Use of these functions results from a circular mask; rectangular masks would use Legendre polynomials. We speculate that a few other mask shapes would have similarly natural basis functions. However, the practical gains which we have found result from the circular mask are tremendous.

They begin with the properties of the Bessel functions of high parameter, which give us a natural means of selecting the end of the set of basis functions. The number of such required functions for describing an apodization scale as the square of the mask size. It can easily be below 100 for small masks, even for complex pupils. The mask size itself gives us a scaled area of the pupil, which we have shown is exactly equal to the sum of the eigenvalues of the matrix — a very convenient error check.

The reduction of size is considerable compared to current pupil discretized descriptions, which can require matrices of size 2000×2000 (Krist et al., 2011), albeit for full end-to-end propagation with distortions. This is somewhat offset by the necessity of calculating a numerical integral over the pupil for

each unique pair of basis functions to produce the kernel elements.

The next major benefit is our ability to use the Bessel recursion functions. These let us write the product of pupil-plane polynomials r^n on the apodization as a linear operator on our basis. While this will require more basis functions than might be needed for the pupil and mask at hand, the result is that we can explicitly include pupil-scale aberrations as perturbations with minimal difficulty. The basis modes necessary to accommodate a given perturbation for a given mask size are also a fixed quantity, not dependent on the pupil geometry.

Based on the rates of convergence, we believe that r^5 or r^6 will be handled without difficulty, sufficient to manage low-order aberrations. We expect this to be of great use, reducing working times for testing their effects when incorporated into pre-existing frameworks (Krist et al., 2011) (Laurent et al., 2018) (Leboulleux et al., 2018). Speculatively, if an aberration probability distributions could be written then this linear treatment could allow direct integration over the state space to produce estimators.

We can also write the action of the mask as a linear operator. This will be generally true, as changes to fields on the mask do not produce fields off the mask, and so are still sums over basis functions on the mask. It is relatively simple to write the result of acting on the Zernike polynomials with functions of even power in ρ or described as a sum over other Zernike polynomials. We have also shown that the action of the pure phase operator $e^{i\ell\theta}$ can still be described easily in our basis if we only care about the resulting fields at the Lyot stop.

Another unexpected benefit of the Bessel functions is that the operator $r\partial_r$ exists as a simple and well-defined matrix. This operator leads to the change of scale or dilation operator, which we have shown is directly related to the change of wavelength operator. While we do not have an explicit formula for the matrix entries, they are simple polynomial functions of the dilation factor and can be calculated explicitly by appropriate software (e.g. Mathematica™).

Consequentially, we have been able to relate the kernels for the coronagraph at any two wavelengths using a simple linear transformation. We will be able to very cheaply produce apodizations at different wavelengths, as well as their eigenvalues. We expect this to be of great benefit given the current 20% broadband goal for terrestrial exoplanet detection (Krist et al., 2011).

It is best if $\eta \equiv \lambda_{old}/\lambda_{new} < 1$ for this transformation. Errors accumulate rapidly for the other case, due to reliance on truncated basis modes. We use 1.1 as a practical upper bound, as this did not push the radial mode cutoff past 30. Going too far in the other direction can produce large relative errors in eigenvalues, if the eigenvalues themselves drop past 10^{-10}; however, we were able to reproduce eigenvalues to within a relative 10^{-5} with $\eta = 0.2$, far beyond bandwidth requirements.

The same math for the dilation operator also has allowed us to write a general proof that for all pupils, the eigenvalues must increase with increasing $\mathcal{N}$ (decreasing λ).

While not a direct consequence of the Bessel functions, there are additional benefits from considering propagation of light from the pupil to the Lyot stop in terms of the Slepian modes. The blank pupil, itself, can be written as a sum

over these modes, allowing analysis of non-apodized coronagraphs within this general framework. Expressions for the Lyot-plane residual energy and the pupil photon throughput have natural simple forms. Using these forms, we are in fact able to state that the throughput will always be a weighted sum of the different mode throughputs, placing strong bounds on our capability to improve that metric. We can also show that the APLC and phase mask residual energies are likewise weighted sums in the case where the Lyot stop is the same form as the pupil.

Propagation of wavelengths other than the design wavelength can be incorporated using the same dilation operator as before (so long as the elements of the coronagraph do not themselves cause chromatic aberration). In this case, however, it is done using $(\eta)^{-1}$ as the parameter. If we wish to examine both the eigenvalues over the bandwidth and use the dilation operator to study the propagation, we have two options. The first is to produce the kernel for the center of the bandwidth, and take care to use $1.1^{-1} \approx 0.9 \leq \eta \leq 1.1$. The alternative is to calculate the kernel for both ends of the bandwidth, using the appropriate one for eigenvalue calculations and the other for propagation. The latter approach allows for an error check by comparing the dilated eigenvalues.

This propagation will require more careful handling around phase-altering coronagraphic structures. Rather than direct use of the dilation matrix, the phase effects will need explicit functional form in η. With this, however, the dilation matrix can be used for the propagation. We believe that this will offer continued benefits for analysis once integrated with current methods.

One difficult faced by this formalism is the handling of off-axis sources. Very small deviations can be handled by simple Taylor expansion, producing polynomials in r which we have already discussed. While somewhat larger deviations can be incorporated, the approximation becomes increasingly error-prone as it must rely on $1/\Lambda_a$ factors and due to the very large number of angular modes required. We do not believe that we can reliably treat off-axis sources of light past a couple λ/D_P. This is sufficient to describe pointing errors and finite star effects.

Having explored the abstract application of this formalism to the coronagraphic problem, we then used it replicate some prior results from (Soummer et al., 2009). As it did so with flying colors, our confidence in our methodology is much improved. We then extended our analysis of the apodizations for circular, centrally-obstructed pupils. We were able to show that the eigenvalues obey a strict ordering caused by the symmetry of the pupil. The eigenvalues also displayed a curious behavior, in that the odd-angular-mode eigenvalues were at their most sensitive chromatic position when the even-angular-modes were at their most achromatic, and vice-versa. We unfortunately lack a reason for this behavior.

The eigenvalues for any set $\mathcal{N}$ and $\mathcal{R}_S$ were shown to be well-aligned with a simple function $[1 + e^{c\left(\frac{a-\mathrm{Tr}K}{\mathrm{Tr}K}\right)}]$ which matches to the form proven for the prolate spheroidals in (Slepian, 1964). While not a least-squares fit, we found that the constant $c = \mathcal{N}(1 - R_S)$ acceptably well. Intriguingly, this formula implies that at $\Lambda_a = 1/2$, $\frac{\partial\Lambda}{\partial\mathcal{N}}$ is at its largest, a fact which is true for any $c \propto \mathcal{N}$. The pattern of eigenvalues for the irregular JWST pupil conformed

to this pattern, which implies that phase masks are at the chromatically most sensitive position possible for the eigenvalues.

We also addressed the "bell-bagel" transition observed in circular pupils (Soummer et al., 2009), a qualitative change of the apodization mode found by previous algorithms. We were able to show that this occurs precisely when two elements of the eigenvectors attain the specific value $1/\sqrt{3}$. We have a general argument that this occurs when changing parameters alter the ratio of the power those two modes have within the pupil, though we have not pursued this quantitatively.

This argument is in line with other observed localization transitions. As $\mathcal{N}$ is increased, Bessel functions of higher order are of use inside the pupil. Once the necessary order to allow angular modes of width comparable to pupil features enters, localization is possible.

We then turned to a non-circular hypothetical APLC using the Webb primary mirror as the pupil, including segmentation and support members. As predicted, our methodology was able to produce the apodization family, requiring as few as 55 basis functions to do so tolerably well for this proof of concept. For these modes we demonstrated the creation of the instrument-plane PSFs, using a numerical FFT from the Lyot plane. As an aside, we showed that such an FFT would necessarily be restricted to binning of $(1/2)\lambda/D_P$.

We also exhibited angularly-averaged contrast profiles for the top eigenfunctions. They often appeared in pairs or triplets, which gives us reason to believe that linear combinations of apodizations will produce dark zones well. We showed one such linear combination, though did not optimize it for

contrast. Instead, we showed that we can produce a wholly positive mode by mimicking the circular pupil's positive mode. Doing so required a sum of only three of the Slepian functions, providing a good starting point for optimization codes. These can use the coefficients as the variables, reducing the space to be searched.

The non-circular pupil demonstrated the troubles that occur when different eigenmodes reach degeneracy (equal eigenvalues). On some occasions, the mode vanished as a separate recognizable entity, with new modes (linear combinations of old modes) taking their places. In other situations, while we were able to identify individual modes across the degeneracy point, their relative ranking changed. This produces spurious errors when following eigenvalues and their derivatives, and will need to be controlled for in future work.

Extensions

We can see many ways in which this work can be extended. The most obvious, as we have alluded to, is to determine the suitability for and integrate into pre-existing analysis codes ((Ruane et al., 2018) and many more). Whether or not finite-mask approximations for the infinite-mask styles of band-limited and vortex coronagraphs are within acceptable error will have to be determined.

We have also developed a general formula for angularly-averaged intensity in the instrument plane as a function of radial distance. While this is nothing more than the familiar Fourier transform in new terms for the arbitrary pupil, we were able to show that significant simplification occurs for the circular

pupil. We have not attempted to explore any consequences of our general expression, believing it to be of no practical use given the repeated numeric integrations involved. It is possible that this is still a useful expression; angular-sector segments produce analytical results on integration, which may allow for continued analysis of their effects through simplification of the general case.

We have speculated that an end-to-end Slepian problem may be feasible. If the working region of the instrument plane is annular, then the annular Zernike polynomials would serve as a basis. The feasibility of a Slepian dual approach, as with our optical reversal, would need to be addressed. Alternatively, we could construct a Lyot stop-working region Slepian system. Either scenario would intend to work with low-eigenvalue Slepians, as those are outside the region, and so would need to take additional caution.

Our study of the basis functions and their mode numbers related to the behavior of the vortex coronagraph. We therefore have a large selection of functions whose behavior is to produce a large dark hole, in direct contrast to those basis functions whose Fourier transform is a limited region of light. The action of these functions once restricted to a pupil or the Lyot stop could be investigated as an alternate means of diverting starlight.

Non-circular masks were almost entirely neglected in this study, as circular masks had such a wide range of benefits. It is possible that we have overlooked similar benefits for these alternative shapes. If that is the case, then the machinery we have developed here can be deployed for their use.

We encourage any reader to take these possibilities, or any other direction

we have overlooked, and continue to extend the work we have done here.

References

Aime, C., R. Soummer, and A. Ferrari (2002). "Total Coronagraphic Extinction of Rectangular Apertures Using Linear Prolate Apodizations". In: *Astronomy and Astrophysics* 389.1, pp. 334–344.

Soummer, R. (2005). "Apodized Pupil Lyot Coronagraphs for Arbitrary Telescope Apertures". In: *The Astrophysical Journal* 618.2, pp. L161–L164.

Slepian, D. and H. O. Pollak (1961). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — I". In: *The Bell System Technical Journal* 40.1, pp. 43–63.

Slepian, D. (1964). "Prolate Spheroidal Wave Functions, Fourier Analysis, and Uncertainty — IV: Extensions to Many Dimensions and Generalized Prolate Spheroidal Functions". In: *The Bell System Technical Journal* 43.6, pp. 3009–3057.

Slepian, D. (1976). "On Bandwidth". In: *Proceedings of the IEEE* 64.3, pp. 292–300.

Soummer, R., L. Pueyo, A. Ferrari, C. Aime, and A. Sivaramakrishnan (2009). "Apodized Pupil Lyot Coronagraphs for Arbitrary Apertures, II. Theoretical Properties and Application to Extremely Large Telescopes". In: *The Astrophysical Journal* 695.1, pp. 695–706.

Krist, John E., Ruslan Belikov, Laurent Pueyo, Dimitri P. Mawet, Dwight Moody, John T. Trauger, and Stuart B. Shaklan (2011). *Assessing the performance limits of internal coronagraphs through end-to-end modeling: a NASA TDEM study*. DOI: 10.1117/12.892772. URL: https://doi.org/10.1117/12.892772.

Laurent, Kathryn St., Kevin Fogarty, Neil T. Zimmerman, Mamadou N'Diaye, Christopher C. Stark, Johan Mazoyer, Anand Sivaramakrishnan, Laurent Pueyo, Stuart Shaklan, Robert Vanderbei, and Rèmi Soummer (2018). *Apodized pupil Lyot coronagraphs designs for future segmented space telescopes*. DOI: 10.1117/12.2313902. URL: https://doi.org/10.1117/12.2313902.

Leboulleux, Lucie, Laurent Pueyo, Jean-François Sauvage, Thierry Fusco, Johan Mazoyer, Anand Sivaramakrishnan, Mamadou N'Diaye, and Rèmi Soummer (2018). *Sensitivity analysis for high-contrast imaging with segmented space telescopes*. DOI: 10.1117/12.2313904. URL: https://doi.org/10.1117/12.2313904.

Ruane, G., A. Riggs, C. T. Coker, S. B. Shaklan, E. Sidick, D. Mawet, J. Jewell, K. Balasubramanian, and C. C. Stark (2018). *Fast linearized coronagraph optimizer (FALCO) IV: coronagraph design survey for obstructed and segmented apertures*. DOI: 10.1117/12.2312973. URL: https://doi.org/10.1117/12.2312973.

Chapter 7

Curriculum Vitae

David Ely

2615 Maryland Avenue #3
Baltimore MD 21218
M +1 (804) 239 3259
E d.ely@alum.mit.edu

Education

Ph.D., *Johns Hopkins University*, Baltimore MD. **2009–2018**

S.B., *MIT*, Cambridge MA. **2004–2008**

Doctoral thesis

Title: *A Unified Framework for Finite-Mask Coronagraphy as Applied to Exoplanet Imaging*

Supervisors: Laurent Pueyo, Julian Krolik

Description: Expansion and application of band-limited eigenfunctions for the efficient design of and the propagation of light through general coronagraphs.

Experience

Vocational

Teaching Assistant, *Johns Hopkins University*, Baltimore MD. **2009–2018**

Homework and test design and grading
Recitation, tutoring, and office hours
Classes ranging from introductory mechanics to graduate quantum field theory
Detailed achievements:

- EJ Rhee Award for excellence in teaching;
- Co-designed and equipped a new "hands-on" undergraduate mechanics course;
- Head TA for undergraduate mechanics and electromagnetism; organized and oversaw 5-10 other TAs while continuing own TA duties.

Research Scientist, *InnovX*, Woburn, MA. **2008–2009**

Algorithmic design and implementation for metal-content analysis through X-ray fluorescence.

Languages

Latin: Moderate, but very rusty *Gold medal, National Latin Exam*

Russian: Some low-level reading and listening *Self-taught*

Computer skills

Office Software: LibreOffice and MSOffice document, spreadsheet, and powerpoint

Programming: Object-oriented C++; LaTeX, some Python

Operating Systems: Windows 7; Linux/Ubuntu

Hardware: Simple component assembly, soldering, troubleshooting

Interests

Reading: Primarily science fiction, history, technology

Cooking: Simple meals for home

References

Available on request